НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

На правах рукопису
УДК 004.4

До захисту допущено
В. о. завідувача кафедри ММСА
О.Л.Тимощук
«___» _____ 2019 р.

Магістерська дисертація

на здобуття ступеня магістра за спеціальністю 124 Системний аналіз
на тему: «Система оцінювання кредитоспроможності позичальників з використанням методів інтелектуального аналізу даних»

Виконав:
студент ІІ курсу, групи КА-з81 мп
Карсунцева Єлизавета Вадимівна _____

Керівник: доцент кафедри ММСА
д.т.н, доц. Кузнєцова Н.В. _____

Рецензент:

_____

Засвідчую, що у цій магістерській дисертації
немає запозичень з праць інших авторів
без відповідних посилань
Студент _____

Київ
2019

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

Рівень вищої освіти — другий (магістерський)
Спеціальність — 124 «Системний аналіз»

ЗАТВЕРДЖУЮ
В. о. завідувача кафедри ММСА
О. Л. Тимощук
«___» _____ 2019 р.

ЗАВДАННЯ
на магістерську дисертацію студента Карсунцевої Єлизавети Вадимівни

1. Тема дисертації: «Система оцінювання кредитоспроможності позичальника з використанням методів інтелектуального аналізу даних», науковий керівник дисертації Карсунцева Єлизавета Вадимівна, к.ф.-м.н., доцент, затверджені наказом по університету від «05» листопада_ № _3825-с__

2. Термін подання студентом дисертації: 13 грудня 2019 р.

3. Об'єкт дослідження: Кредитоспроможність позичальника, її сутність, методи та підходи оцінювання

4. Предмет дослідження: методи математичного моделювання, які застосовуються в задачах оцінювання кредитоспроможності

5. Перелік завдань, які потрібно розробити:

1) дослідити сучасний стан та особливості застосування математичного моделювання для оцінювання кредитоспроможності позичальника;

2) розробити математичну модель за допомогою логістичної регресії (метод максимальної правдоподібності з використанням методу градієнтного спуску);

3) розв'язати розроблену математичну модель та на її основі створити програмний продукт;

4) пошук даних для застосування в програмі;

5) розробити стартап-проект виведення на ринок результатів дослідження;

6)розробити концептуальні висновки за результатами наукового дослідження

6. Орієнтовний перелік графічного (ілюстративного) матеріалу:

1)Схема системи аналізу кредитоспроможності клієнтів у банку;

2)Схема побудованої скорингової моделі за допомогою дерев рішень;

3)Приклади функціонування створеного програмного продукту;

4)Таблиці у розділі стартап-проекту.

7. Орієнтовний перелік публікацій:

Публікація наукової статті у фаховому журналі «Системні дослідження та інформаційні технології»: Розробка моделі оцінювання кредитоспроможності позичальника з використанням методів інтелектуального аналізу даних

8. Дата видачі завдання: 05 вересня 2019 р.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації
1.	Концептуальний вступ дисертації. Формулювання об'єкта, предмета, цілі, завдань, новизни, практичної значущості результатів	05.09.2019—20.09.2019
2.	Перший розділ. Огляд літературно-інформаційних джерел. Понятійно-категоріальний апарат. Характеристика об'єкта	21.09.2019—30.09.2019
3.	Другий розділ. Розробка математичної моделі для задачі оцінювання кредитних ризиків (підготовчий аналіз та обробка даних для розробки скорингової моделі). Формування навчальної та тестової вибірки.	31.09.2019—16.10.2019
4.	Третій розділ. Огляд методів та підходів щодо оцінювання скорингових моделей (ROC- крива та індекс Джині, статистика Колмогорова-Смірнова)	17.10.2019—29.10.2019
5.	Четвертий розділ. Імплементація отриманих результатів у програмний продукт. Тестування програми	27.10.2019—02.11.2019
6.	П'ятий розділ. Стартап-проект	03.11.2019—15.11.2019
7.	Концептуальні висновки. Перспективи розвитку отриманих рішень	16.11.2019—30.11.2019

Студент Є.В.Карсунцева

Науковий керівник дисертації Н.В. Кузнєцова

РЕФЕРАТ

Магістерська дисертація: 96с., 30 рис., 30 табл., 2 додатки, 23 джерел.

Актуальність теми: Кредитний ризик банківської установи як один з видів банківських ризиків є головним об'єктом уваги фінансово-кредитних установ. Кредитна політика банків має обов'язково враховувати ці ризики, запобігати їх виникненню та кваліфіковано ними управляти, тобто зводити до мінімуму можливі негативні наслідки проведення кредитних операцій. У зв’язку з нинішнім кризовим станом у банківській сфері постає нагальним застосування та розробка нових більш досконалих методів оцінювання кредитних ризиків і кредитоспроможності осіб.

Мета даної роботи полягає у дослідженні та вдосконаленні існуючих методик побудови скорингових моделей та розробці системи підтримки прийняття рішень для оцінювання кредитоспроможності фізичних осіб з використанням методу логістичної регресії.

Об’єкт дослідження: база даних з аплікаційними характеристиками клієнтів. Предмет дослідження: моделі і методи оцінювання кредитоспроможності позичальників. Методи дослідження: метод логістичної регресії, метод максимальної правдоподібності, метод градієнтного спуску.

Програмний продукт реалізований за допомогою мови програмування C# у середовищі розробки Microsoft Visual Studio 2012. Для порівняльного аналізу отриманих результатів були побудовані моделі у вигляді дерев рішень і скорингової карти в системі SAS Enterprise Miner.

Отримані результати: розроблено систему підтримки прийняття рішень для прогнозування кредитоспроможності фізичних осіб з використанням методу логістичної регресії та методу максимальної правдоподібності.

КРЕДИТОСПРОМОЖНІСТЬ, КРЕДИТНИЙ СКОРИНГ, ЛОГІСТИЧНА РЕГРЕСІЯ, ТОЧНІСТЬ МОДЕЛІ, ІНДЕКС GINI

ABSTRACT

Theme: “The borrower's credit rating system using data mining techniques”.

Master's thesis explanatory note: 96 p., 30 fig., 30 tab., 2 appendices, 23 sources.

Actuality: Credit risk of a banking institution as a type of banking risk is a major focus of attention of financial institutions. The credit policy of banks must take into account these risks, prevent them from occurring and be qualified to manage them, ie minimize the possible negative effects of credit operations. Due to the current crisis in the banking sector, it is urgent to apply and develop more sophisticated methods for assessing credit risk and creditworthiness of individuals.

The purpose of this work is to study and improve existing methods of constructing scoring models and to develop a decision support system for assessing the creditworthiness of individuals using the method of logistic regression.

Object of study: a database with customer characteristics. Subject of research: Models and methods of assessing the creditworthiness of individuals. Research methods: logistic regression method, maximum likelihood method, gradient descent method.

The software product was implemented using the C# programming language in the Microsoft Visual Studio 2012 development environment. For a comparative analysis of the results were built models as decision trees and scorecard in the SAS Enterprise Miner system.

Obtained results: a decision support system was developed for predicting the creditworthiness of individuals using the logistic regression method and the maximum likelihood method.

CREDITWORTHINESS, CREDIT SCORING, LOGISTIC REGRESSION, ACCURACY OF THE MODEL, INDEX GINI

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

CA (англ. Common Accuracy) – загальна точність моделі

IV (англ. Information Value) – інформаційне значення

MAE (англ. Mean Absolute Error) – середня абсолютна похибка

MSE (англ. Mean Squared Error) – середньоквадратична похибка

WOE (англ. Weight of evidence) – зважена сукупність

ПЕОМ – персональна електронно-обчислювальна машина

ПП – програмний продукт

СППР – система підтримки прийняття рішень

ВСТУП

Актуальність теми. В даний час однією з найпомітніших нових тенденцій на ринку банківських послуг є різке зростання інтересу банків до приватних позичальників, а саме розвиток споживчого кредитування. Даний вид кредитування привабливий для банків, адже ставки за споживчими кредитам досить високі (до 60%), що приносить великі прибутки банкам. В даний час існують перспективи розвитку споживчого кредитування, і у виграші залишаться ті учасники ринку, які вчасно встигнуть зайняти лідируючі позиції. В умовах зростання попиту на споживчі кредити загострюється конкуренція між банками. Разом зі зростанням обсягів наданих кредитів збільшуються обсяги простроченої заборгованості, що і збільшує роботу банків з боржниками. Таким чином, конкурентна боротьба на ринку споживчого кредитування йде не просто за позичальника, а за кредитоспроможного позичальника, який адекватно оцінює своє фінансової становище і в змозі повністю погасити борг. Саме тому виникає потреба у застосуванні універсальної прикладної наукової методології системного аналізу у вирішенні системних задач ризик-менеджменту, а саме задач математичного моделювання та прогнозування кредитоспроможності клієнтів банку. Основною сучасною системною методологією прогнозування кредитних ризиків є кредитний скоринг, що полягає у розробці математичних моделей спеціального типу – скорингових моделей, які також називаються скоринговими картами, метою яких є прогнозування майбутнього стану обслуговування позичальником заборгованості, виходячи з соціально-демографічних характеристик, минулих поведінкових індикаторів. Крім того, система кредитного скорингу дозволяє швидко впоратись з великим обсягом кредитних заявок, знижуючи експлуатаційні витрати та мінімізуючи затрати через помилки недосвідчених співробітників банку.

Отже, метою даного дослідження є вдосконалення системної методології побудови моделей оцінювання кредитоспроможності, розробка та вдосконалення методів і алгоритмів обчислення ключових показників, методів моделювання, розробка оригінального програмного продукту. Для досягнення мети потрібно вирішити такі завдання:

1)провести огляд та аналіз існуючих математичних методів моделювання і прогнозування кредитних ризиків;

2)розробити систему підтримки прийняття рішень для аналізу, моделювання та прогнозування ймовірності повернення наданого споживчого кредиту;

3) розробити методом розрахунку статистики Колмогорова-Смирнова і ваг категорій змінних та інформаційної статистики за умови відомого розподілу категорій та умовному розподілі цільової змінної.

Об’єкт дослідження: база даних з аплікаційними характеристиками клієнтів. Предмет дослідження: моделі і методи оцінювання кредитоспроможності фізичних осіб. Методи дослідження: метод логістичної регресії, метод максимальної правдоподібності, метод градієнтного спуску.

Отримані результати: розроблено систему підтримки прийняття рішень для прогнозування кредитоспроможності фізичних осіб з використанням методу логістичної регресії та методу максимальної правдоподібності.

Робота складається з 4 розділів. В першому розділі розглядаються поняття кредитоспроможності та кредитного скорингу, проводиться огляд програмного забезпечення, призначеного для банківської клієнтської аналітики. У другому розділі вивчаються методи побудови скорингових систем та процес попереднього аналізу і обробки даних. У третьому розділі дисертації описується розроблена СППР, надається інструкція з експлуатації програмного продукту, а також проводиться порівняння результатів роботи системи з іншими методами. Четвертий розділ присвячено розробленню стартап-проекту на основі створеного програмного продукту.

1. ДОСЛІДЖЕННЯ ПРЕДМЕТНОЇ ОБЛАСТІ

1.1 Поняття кредитоспроможності та скорингу

Основним критерієм, що формує кредитні відносини між банківською установою та потенційним клієнтом, є кредитоспроможність позичальника. Саме визначена банком кредитоспроможність позичальника є необхідною умовою для укладення кредитного договору і дає можливість визначити фактори, які впливатимуть на невиплату кредиту.

Кредитоспроможність – наявність у позичальника передумов для видачі кредиту і його здатність повернути борг в обумовлені договором строки та у повному обсязі.

Кредитоспроможність позичальника визначається за такими його характеристиками як:

а) здатність своєчасно розраховуватися за раніше одержаними кредитами;

б) поточне фінансове становище;

в) спроможність мобілізації коштів з інших джерел.

Скоринг формує ряд показників, аналіз яких є корисним при створенні методики оцінки кредитоспроможності позичальників банку. Скоринг дозволяє приймати рішення про видачу кредиту всього за кілька секунд. Але ухваленню рішення передує збір інформації про клієнта. Найбільш істотним є визначити, які характеристики фізичної особи-клієнта банку мають фінансовий сенс і відображають його здатність та бажання своєчасно погашати кредит. За своєю суттю система покликана категоризувати оцінку ступеня кредитного ризику по потенційному позичальникові. У найпростішому і найбільш вагомому для практики випадку ця оцінка бінарних: «видати кредит» або «відмовити у видачі кредиту». Основною метою і завданням кредитного скорингу з точки зору банківської діяльності є управління і мінімізація кредитного ризику за рахунок якісного відбору позичальників.

Скоринг виник перш за все як метод аналізу кредитоспроможності юридичних осіб в силу різкого скорочення професійних кредитних інспекторів в банках в ході Другої світової війни. Класиками в розробці скорингових моделей можна вважати Е. Альтмана, Дж. Фулмера, Р. Чессера, крім того велике значення мали дослідження компанії Fair Isaac, метою таких моделей було і залишається визначити ймовірність банкрутства позичальника або ймовірність дефолту (від англ. default - несплата, невиконання). У США 1967 року для оцінювання кредитоспроможності за допомогою скорингу вперше було застосовано інформаційні технології, що дозволило скоротити частку безнадійних кредитів на 50%. У 1980-х рр. було запропоновано модель скорингу на основі нейромереж, що підвищило прибутковість існуючої моделі на 27%.

Таким чином, кредитний скоринг (від англ. score — бал) — це система оцінки кредитоспроможності банками, що на основі кредитної історії оцінює ймовірність дефолту потенційного позичальника, виходячи з його соціально-демографічних характеристик. Маючи базу даних поганих і гарних кредитів, фінустанова за допомогою статистичних інструментів може виявити фактори, що впливають на здатність і бажання клієнта повернути борг. Загально прийнято, що існує кореляція між певними соціальними даними і надійністю позичальника [1].

Існує чотири види кредитного скорингу – аплікаційний, поведінковий, колекторський і шахрайства. Аплікаційний використовується на початку кредитних відносин, коли позичальник отримує кредит. Поведінковий є частиною кредитного моніторингу. Колекторський застосовують у разі невиконання позичальником своїх зобов’язань. Але українські банки найчастіше використовують лише аплікаційний скоринг, здебільшого це пояснюється нерозвиненістю системи скорингу в Україні та високими цінами на послуги розробників скорингових моделей.

Як зазначають фахівці, у типовій скоринговій моделі – від 13 до 25 параметрів: 13 – для споживчого кредитування і 25 – для автокредитування чи

іпотеки. Звичайно, для повної та ефективної оцінки кредитоспроможності позичальника 25 параметрів є замало, тобто це говорить про те, що відібрані показники повинні бути найвагомішими при обчисленні кредитоспроможності кожного окремого позичальника [2].

На будь-яких етапах кредитної історії за допомогою скорингу можливе вирішення різноманітних завдань. З огляду на специфіку наявної інформації існує чотири види кредитного скорингу – аплікаційний, поведінковий і колекторський та скоринг оцінки можливості шахрайства.

1.Скоринг за актуальними даними протягом оформлення заявки (application scoring) – оцінка кредитоспроможності потенційних позичальників за наданою інформацією упродовж кредитної трансакції.

2.Скоринг протягом кредитного періоду (behavioral scoring), коли оцінка динаміки стану кредитного рахунку позичальника дозволяє математично оцінити ймовірність повернення кредиту. Скорингові моделі ймовірності, що використовуються для цього, дозволяють спрогнозувати зміну платоспроможності позичальника, визначити оптимальні ліміти за кредитною карткою тощо [14].

3.Оцінка ймовірності повного або часткового повернення кредиту (collection scoring) пропонує визначення пріоритетних напрямів роботи щодо позичальників, коли їхній кредитний рахунок класифікують як «незадовільний». Такий вид використовується за умов порушення позичальником зобов’язань щодо погашення кредиту. Згідно з результатами багатьох досліджень майже 40% неплатежів припадає на позичальників, які невимушено забувають внести платіж за кредитом. Підтримка скоринговою системою collection-скорингу дозволяє автоматично ліквідовувати цю заборгованість.

4.Оцінка можливості шахрайства (fraud scoring) визначає ймовірність потенційних неправомірних дій позичальника. Як правило, цей метод використовується разом з application- і behavioral-скорингом вірогідного аналізу

позичальників. За даними українських банків, під шахрайство підпадає до 10% усіх неплатежів.

Залежно від якості доступної інформації про позичальника і методу прийняття рішень скоринг поділяється на дедуктивний (deductive credit scoring) і емпіричний (empirical credit scoring) [16].

Підґрунтям дедуктивних систем є висновки та оцінки експертів, відповідно до критичних значень і вагомості оцінок окремих критеріїв. Порівняно з емпіричним скорингом, ці системи вважаються менш адекватними з огляду на людський фактор і ризик суб'єктивної оцінки експерта. У свою чергу, основними елементами емпіричних систем скорингу є статистичні моделі та методи ранжирування неоднорідних багатовимірних даних. При цьому вибір методу класифікації значною мірою залежить від категорії позичальника і виду кредиту. Це зумовлено значними відмінностями статистичних моделей з огляду на відповідні види кредитування: іпотечне, споживче, на придбання авто тощо.

Характерною особливістю банківського ринку трансформаційних країн є відсутність відповідної статистичної бази щодо функціонування грошово-кредитних систем, що значно утруднює створення адекватних систем скорингу. Перш за все, це стосується application-скорингу, зважаючи на необхідність доступу до бази даних кредитних справ у цілому по системі.

Системи скорингу, побудовані на підґрунті запропонованого алгоритму, надають здатність ефективно та адекватно оцінити кредитоспроможність позичальника банку в умовах кризи і посткризової рецесії, оскільки враховують регіональну специфіку, загальносвітові глобалізаційно-інтеграційні процеси та динаміку змін макроекономічної ситуації [12].

Перевагою рейтингової моделі є те, що вона проста у використанні, дозволяє розрахувати оптимальні значення на основі кількох показників. Однак при використанні цієї методики слід ураховувати необхідність ретельного відбору фінансових показників. Використовуються показники, які відображають різні напрями роботи позичальника: необхідність обґрунтування

нормативних значень показників; визначення величини відхилення у граничних зонах, що дозволяє відносити позичальників до різних класів; при рейтинговій оцінці однакові показники використовуються тільки до оптимальних значень, які відповідають установленим нормативам, але не враховується рівень їх виконання чи невиконання; фінансові коефіцієнти відображають попередній період на основі даних про залишки; розраховані коефіцієнти відображають тільки окремі напрями діяльності позичальника; у системі коефіцієнтів не враховується велика кількість факторів, зокрема репутація позичальника, перспективи та особливості ринкової кон'юнктури, оцінка товарної продукції, перспективи капіталовкладень тощо.

Fraud-scoring – перспективний напрям кредитного скорингу. Ризики зростання портфеля неповернень змусили фінінститути всерйоз задуматися про методи боротьби з такими позичальниками. Одним із них став скоринг із визначення шахрайства (fraud-scoring), що дозволяє банку в онлайн-режимі виявляти здобувачів, чиї звернення варто відхилити або відкласти для більш детального розгляду. Скорингові моделі для виявлення спроб обдурити фінустанову підрозподіляють усіх потенційних позичальників на групи за ймовірністю, що те чи інше прохання про видачу кредиту є шахрайством. Fraud-scoring – новий, але перспективний напрям. Його вже розробляють деякі вітчизняні банки, зокрема, ПАТ «UnicredutBank». Після сплеску неповернень ПАТ «Дельта Банк» збільшив кількість питань до 105-ти [16].

Такі заходи не тільки надають додаткову інформацію для поведінкового і колекторського скорингу, а й можуть заплутати шахрая, оскільки скорингова система враховує не всі питання. Як правило, банки обмежуються придбанням скорингової карти для оцінювання платоспроможності позичальника і рідко купують у того ж розроблювача програми для служб зі збору боргів. Розробників скорингових рішень фінустанови обирають, виходячи з цін на їхню продукцію і досвіду в цій сфері. У нас такі програми пропонують дві компанії: вітчизняна «Скорто Солюшенс» і міжнародна фірма «SAS», що працює з банками в усьому світі.

Основні вимоги до скорингових моделей. Зазначимо, що зовсім не обов'язково займатися створенням скорингових моделей своїми силами. Існує досить компаній, здатних виконувати вказану роботу за кредитну установу. Лідером на цьому ринку вважається компанія «Faim Isaac Corporation». Саме її системи використовуються в більшості провідних американських банків. Також слід відмітити, що за допомогою скорингової моделі можна прогнозувати кредитний ризик лише на певний період часу, що, наприклад, на Заході триває від одного до двох років. Тобто релевантність моделі з часом має тенденцію до зниження. Що стосується України, то внаслідок відомих причин він не може становити більше ніж рік. Таким чином, налагодивши модель, необхідно її доопрацьовувати.

Успішність скорингової моделі пояснюють деякі ключові фактори: неупередженість оцінки (скоринг відмежовує суб'єктивність оцінок, традиційно пов'язану з кредитними рішеннями); стандартизація кредитних оцінок; можливість автоматизації; контроль (завдяки стандартизації кредитних банкам не важко контролювати і відстежувати ефективність кредитних рішень); зростання дохідності (автоматизація знижує витрати на ручне опрацювання заявок на кредит до мінімуму) [7].

Визнаючи безсумнівні переваги скорингового кредитування, закордонні банки вкладають у його розробку великі зусилля, не шкодуючи ні коштів, ні часу. Однак скорингова система аналізу кредитних заявок повинна бути статистично вивірена і вимагає високого професіоналізму та постійного поновлення інформації і вдосконалення моделей. Однією з найважливіших функцій є створення скорингових моделей. Особливо важливо, щоб була можливість централізованої інтеграції в роботу новостворених моделей у найкоротші терміни. Для цих цілей використовуються спеціалізовані програми, які дозволяють створювати різні моделі оцінки позичальників, починаючи від простих бальних і закінчуючи кластерним аналізом, деревами рішень і нейромережі.

Політика кредитного скорингу на сьогоднішніх, украй динамічних ринках показує, що одних лише скорингових моделей стає недостатньо для прийняття рішення щодо деяких кредитів. «Найкращі» і «найгірші» позичальники видно практично неозброєним оком. Але більша частина кредитних заявок надходить від позичальників, для яких «прямолінійне» застосування скорингових моделей не дасть якісного результату. Виходом із подібної ситуації може стати побудова стратегії прийняття рішень, сегментація і т. п.

У разі, якщо використовуються недостатньо якісні скорингові моделі або вони взагалі відсутні як такі, система кредитного скорингу повинна мати можливість простого створення та управління правилами кредитної політики, тобто створенням систем бонусі/штрафів для оцінки потенційного позичальника. Кредитний аналітик або ризик-менеджер банку, задаючи правила кредитної політики, отримує можливість уточнити систему формування рейтингу позичальника, тобто одні правила й умови можуть збільшувати рейтинг, інші – зменшувати і т. п.

Коли рішення скорингової системи за кредитною заявкою неоднозначне, то на перший план виходить можливість створення та управління правилами розподілу заявок для кредитних фахівців з урахуванням їхніх прав і повноважень. Тобто визначається, якого фахівця, до якого рівня повинна відправлятися та чи інша заявка на розгляд.

«Рівнів» фахівців може бути багато, починаючи зі співробітника економічної безпеки і закінчуючи кредитними експертами. Причому завдання подібних правил розподілу заявок має бути максимально простим і доступним для ризик-менеджерів та кредитних аналітиків і не вимагати спеціальних знань.

Також із важливих елементів підтримки процесу прийняття рішення є можливість гнучкого налаштування інтерпретації скорингового рейтингу для кредитних фахівців. Якщо така можливість реалізована – скорингова система може видати для кредитних фахівців рекомендації, зауваження, підказки і різного роду повідомлення, роблячи, таким чином, оцінку позичальника максимально об’єктивною і якісною. Правила формування подібного роду

повідомлень визначає кредитний департамент або департамент ризик-менеджменту [5].

Украй важливо щоб скорингова система передбачала можливість швидкої і якісної оцінки динаміки зміни як стану кредитного рахунку окремого позичальника, так і кредитного портфеля в цілому. Тобто повинна бути реалізована система скорингової звітності, на підставі якої можна відстежувати адекватність роботи всієї системи кредитного скорингу, використовуваних скорингових моделей і стратегій оцінки позичальників [10].

Переваги скорингових систем. Як свідчить досвід західних країн, після введення в роботу скорингових моделей рівень «поганих» боргів скоротився на 15– 20% у порівнянні з ручним опрацюванням кредитних заявок.

Проте досвід російських фінансистів не такий оптимістичний. За різними оцінками, відсоток «неповернень» становить від 10 до 20%. Метод скорингу дозволяє провести експрес-аналіз заявки на кредит у присутності клієнта.

У французьких банках клієнт, запросивши позику і заповнивши спеціальну анкету, може отримати відповідь про можливість надання позики протягом кількох хвилин.

Серед переваг скорингових систем західні банкіри вказують, перш за все, на зниження рівня неповернення кредиту. Далі наголошується швидкість і безсторонність в ухваленні рішень, можливість ефективного управління кредитним портфелем, відсутність необхідності тривалого навчання персоналу [12].

У практиці більшості американських банків для оцінки позичальника використовують «правило п’яти сі»:

-1С (customer’s character – характер позичальника) – репутація позичальника, ступінь відповідальності, готовність і бажання сплатити борг;

-2С (capacity to pay – фінансові можливості) – припускає ретельний аналіз доходів і витрат позичальника і перспективи їх- нього розвитку в майбутньому;

- 3С (capital) – капітал, майно;

- 4C (collateral) – забезпечення позики, достатність, якість і ступінь реалізовуваної застави в разі непогашений позички;

- 5C (current business conditions and goodwill – загальні економічні умови)– визначають діловий клімат у країні і впливають на становище банку і позичальника.

Перераховані критерії «сі» іноді доповнюють шостим критерієм – 6 C (control) – моніторинг законодавчих основ діяльності позичальника і відповідність його стандартам банку. Цікавим є досвід зарубіжного банківського сектору із залучення до оцінки кредитного ризику незалежних рейтингових агентств [6]. Рейтингове агентство має у своєму розпорядженні великий обсяг інформації і досвід створення неупереджених оцінок для всіх можливих варіантів ситуацій, у нього відсутня будь-яка зацікавленість, крім формування достовірної оцінки кредитного ризику банку.

Скорингова система складається з вхідного потоку, скорингової моделі та вихідного потоку (рис.1.2). Ключовим елементом скорингової системи є скорингова модель. Це економіко-математична модель, найчастіше реалізована у програмному пакеті, або у програмному коді як окрема програма, прив’язана до бази даних, що кожному позичальнику ставить у відповідність число (скоринговий бал), якому відповідає ймовірність настання кредитної події.

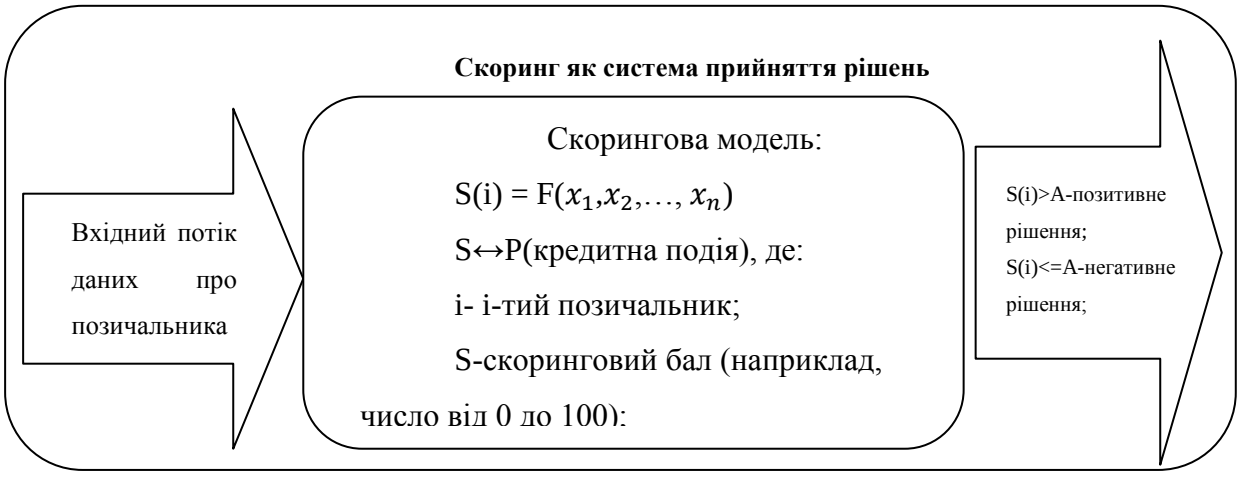

Рисунок. 1.2 — Скорингова система

У світовій практиці досить широко застосовуються такі системи аналізу кредитоспроможності як: MEMO RISK (Management – якість менеджменту,

Experience - досвід, Market – загальні обставини для бізнесу позичальника, Operations – оцінка бізнесу позичальника, Repayment – визначення можливості погашення кредиту, Interest – відсоткова ставка, Security - забезпечення, Kontrol — контроль) та система 4 FC (чотири основи кредитоспроможності: Management quality - якість менеджменту, Industry dynamics – специфіка галузі та її динаміка, Security realization – забезпечення та можливість реалізації застави, Financial condition – фінансовий стан позичальника) [16].

Система 4FC (Four Foundations of Creditworthiness – чотири функції кредитоспроможності) досліджує передусім якість управління, що оцінюється фінансовим потенціалом і компетентністю позичальника; динаміку галузі, яка визначається структурними і динамічними умовами галузі, а також конкурентною позицією підприємства; можливість реалізації застави, яка аналізується чистою ліквідною вартістю застави, можливістю контролю заставленого майна; фінансові умови, які засновуються на дослідженні рентабельності, ліквідності та левереджу (фінансового важеля) за певний період часу [11].

Однак було б доцільно використовувати нову систему для оцінювання кредитоспроможності позичальників, ЦЕНЗОР, у кредитній діяльності банків. Він був створений та впроваджений у 1997 році. ЦЕНЗОР складається з таких елементів: Призначення; Експозиція; Наслідки; Програмне забезпечення; Обставина; Репутація (табл. 1.2) [8]. Система CENSOR полегшує підготовку якісних інвестиційних пропозицій, бачення різних аспектів ризику їх проектів, розуміння позичальників та кредиторів, поліпшення якості бізнес-планів та покращення їх оцінки.

Необхідно вказати на наступні особливості запропонованої системи: «ЦЕНЗОР» охоплює всі найважливіші характеристики банку для позичальника, що дає можливість провести всебічний аналіз кредитоспроможності; кожен з елементів нової системи має окреме і чітке призначення, доволі конкретний інструментарій методів оцінювання пов'язаного з ним ризику (що не заперечує їхнього подальшого вдосконалення); жоден із шести аспектів запропонованої

системи аналізу кредитоспроможності не має занадто відмінного від інших навантаження та важливості.

Таблиця 1.2 - Критерії оцінювання фінансового стану позичальника на основі моделей комплексного аналізу

	CAMP	PART	PARS	MEMO	6C	4FC
Репутація позичальника, якість менеджменту	+	-	+	+	+	+
Досвід	-	-	-	+	-	-
Загальні обставини для бізнесу позичальника	-	-	-	+	+	+
Забезпечення кредиту, можливість реалізації застави, спосіб страхування	+	+	+	+	+	+
Контроль	-	-	-	+	-	-
Фінстан позичальника, оцінка бізнесу, адекватність капіталу	-	-	-	+	+	+
Експозиція грошових потоків	+	-	-	-	+	-
Можливість погашення кредиту	+	+	+	+	+	-
Обґрунтування суми кредиту	+	+	+	-	+	-
Процентна ставка	-	-	+	+	-	-
Доцільність надання позики	-	-	+	+	-	-
Мета кредиту	+	+	-	-	-	-
Термін кредиту	-	+	-	-	-	-

Порівняльна характеристика критеріїв оцінювання фінансового стану позичальника на основі моделей комплексного аналізу наведена в таблиці 1.2 [13]. Головним недоліком даних комплексних методик є їх орієнтація здебільшого на якісні чинники, а також треба враховувати той факт, що дані моделі побудовані на основі експертних висновків та в окремих випадках можуть мати суб’єктивний характер.

1.2 Постановка задачі дослідження

Метою магістерської дисертації є дослідження та вдосконалення існуючих методів аналізу і прогнозування кредитних ризиків, розробка програмного забезпечення для попередньої обробки даних і побудови скорингової моделі, та перевірка побудованої моделі на адекватність.

У рамках дисертації необхідно:

- розробити архітектуру системи підтримки прийняття рішень для аналізу, моделювання та прогнозування ймовірності повернення наданого споживчого кредиту;

- розробити програму для побудови скорингової моделі на основі алгоритму логістичної регресії з використанням методу максимальної правдоподібності;

- протестувати комп’ютерну програму на реальних даних та провести порівняльний аналіз з іншими методами.

Для рішення цих задач необхідно дослідити вже існуючі інтелектуальні рішення для вирішення задач керування кредитними ризиками.

Об’єкт дослідження – статистичні дані щодо виданих фінансовими установами споживчих кредитів, які потребують ефективної аналітичної

обробки та є необхідними для побудови скорингових моделей та прийняття рішень при визначені кредитоспроможності клієнтів банку.

Предмет дослідження – математичні методи побудови скорингових моделей, а саме: лінійна регресія, логістична регресія, дерева рішень.

Висновки до розділу 1

Кредитний ризик банківської установи як один з видів банківських ризиків є головним об'єктом уваги фінансово-кредитних установ. Кредитна політика банків має обов'язково враховувати ці ризики, запобігати їх виникненню та кваліфіковано ними управляти, тобто зводити до мінімуму можливі негативні наслідки проведення кредитних операцій. У зв’язку з нинішнім кризовим станом у банківській сфері набуває актуальності застосування та розробка нових більш досконалих методів оцінювання кредитних ризиків і кредитоспроможності осіб.

У даному розділі розглянуто основний інструмент з мінімізації ризику в кредитній діяльності банку, а саме: оцінку кредитоспроможності позичальників. Було дано визначення поняттям кредитоспроможності і кредитного скорингу, а також досліджено основні етапи розвитку скорингу. Розглянуто чотири види кредитного скорингу: application-scoring (аплікаційний скоринг), behavioral-scoring (поведінковий скоринг), collection-scoring (колекторський скоринг) та fraud-scoring (скоринг проти шахраїв).

Показано актуальність та перспективність дослідження, на основі чого сформульовано постановку задачі магістерської дисертації, та виділено етапи її розв’язку.

2 МЕТОДИКА ПРОГНОЗУВАННЯ

Моделі скорингу традиційно використовуються в банківському секторі для оцінки кредитоспроможності позичальників на етапі подання заявок на позику. Оцінка балів дозволяє отримати математично-статистичну модель класифікації спостережень за різними групами відповідно до характеристик цих спостережень.

Різні математичні та статистичні моделі можуть бути використані для побудови бальних систем. Вибір конкретного методу залежить від передумов його застосування та від шкали вимірювання наявної статистики. Це ставить проблему вибору оптимальної моделі, яка б забезпечувала адекватні прогнози для процесів чи систем, що вивчаються. Крім того, дуже важливо підготувати та дослідити дані, що займе 90% вашого часу для створення моделі балів. При побудові моделей для оцінювання кредитного ризику найбільш поширеними ознаками реальної статистики є: формат зібраних даних й наявність пропущених та невірних значень у статистичних вибірках.

Статистичні методи, що лежать в основі скорингових систем, вельми різноманітні. На даний час широко використовуються дискримінантний аналіз, множинна регресія, логістична регресія, дерева класифікації, метод К-найближчих сусідів, байєсовські процедури, метод опорних векторів, MAP-сплайни і нейронні мережі [13].

У цьому розділі аналізуються сучасні підходи та методи побудови бальних систем, кроки в процесі розробки скорингової моделі та шляхи подолання вищезазначених проблем аналізу реальної статистики.

2.1 Методи і моделі для вирішення задачі оцінювання кредитних ризиків

2.1.1 Логістична регресія

Проблема кількісної оцінки та аналізу кредитних ризиків і рейтингів позичальників є актуальною як для закордонних, так і для вітчизняних банків, що кредитують фізичних та юридичних осіб. Методики, які розробляються для кількісної оцінки кредитних ризиків, мають відповідати ряду умов, серед яких особливо важливою є вимога щодо прозорості. Прозорість методики кредитного ризику – це можливість бачити не лише явище в цілому, а й вагомі деталі. Під прозорістю методики розуміють строгість використовуваних математичних методів, згладжування суб'єктивності експертних оцінок, наочність результатів оцінки та аналізу ризику, повне їх сприйняття самими працівниками банків, відкритість методик для контролюючих органів і позичальників. Прозорість методики та результатів досягається шляхом обчислення часток значущих подій (критеріїв) у загальному кредитному ризику. Отже, виникає необхідність застосування такого математичного апарату, що дасть змогу, по-перше, зменшити вплив суб'єктивного чинника при оцінці кредитоспроможності клієнтів та, по-друге, визначити вплив кожного з факторів, що враховуються при аналізі кредитних ризиків на кінцевий результат оцінювання [12].

З метою оцінювання кредитоспроможності позичальника необхідно встановити взаємозв'язок між певним переліком чинників та фактом повернення чи неповернення кредиту позичальником. Повернення та неповернення кредиту може бути позначено за допомогою лише двох значень, зазвичай 0 та 1 (така змінна має назву бінарної). Отже, нам необхідно спрогнозувати значення бінарної змінної. Побудова звичайної множинної регресії в даній задачі не дасть потрібного результату, оскільки розраховані

значення залежної змінної можуть не належати відрізку [0, 1], тому інтерпретація таких результатів ускладнюється. Однак задачу побудови регресійної залежності для такого оцінювання можемо представити не як передбачення значень бінарної змінної, а як моделювання деякої неперервної змінної, яка набуває значення з інтервалу [0, 1]. Такі задачі можуть бути описані за допомогою лінійних моделей ймовірності, або logit- та probit-моделей. Завдяки способу побудови цих моделей прогнозні значення, яких набуває досліджувана змінна, можуть не лише відповідати значенню 0 та 1, а й бути інтерпретованими як ймовірність повернення (неповернення) кредиту позичальником [27].

Розглянемо задачу оцінки кредитоспроможності позичальників – фізичних осіб комерційного банку на основі побудови лінійних моделей імовірності.

Позначимо змінну, яка відповідає стану повернення (неповернення) кредиту, через Bad. Змінна набуває значення 0, якщо кредит повернуто вчасно, та 1 – якщо зобов'язання за кредитом не виконано. Щодо кожного клієнта відома інформація за m-показниками. Вся сукупність інформації в результаті утворює масив X = (xji), (j =1, m; i =1, n). Індекс j визначає номер показника, що обрано для дослідження (j=1, m), i – номер позичальника (i=1, n), n – кількість спостережень у навчальній вибірці.

Класична лінійна модель регресії має такий вигляд:

$$Badi = b_0 + b_1\, x_1 + \ldots + b_m\, x_m, \quad i=1,n, \tag{2.1}$$

де b= (b_0, b_1, …, b_m) – шукані параметри залежності, ui – стохастична складова моделі.

Позначимо P1 = P (Bad = 1) як ймовірність того, що величина Bad набуває значення одиниці. У такому разі модель (2.1) може бути записана у такому вигляді:

$$P1 = P\,(Bad = 1) = X\,b' \qquad (2.2)$$

Модель називається лінійною моделлю ймовірності. Головним недоліком такої моделі є припущення лінійної залежності P1від β. Цей недолік усувають при використанні logit- та probit-моделей. Припускають, що

$$P1 = F(Xb'), \qquad (2.3)$$

де F – деяка функція, область значень якої знаходиться на відрізку [0, 1].

В якості функції F можливо використовувати функцію розподілу деякої випадкової величини. Тоді модель (2) можемо навести, наприклад, у такій інтерпретації:

$$Yi = \beta 0 + \beta 1x1i + \beta 2x2i + \ldots + \beta mxmi + ui, \quad i=1,n, \qquad (2.4)$$

де уі – деяка кількісна змінна, що має лінійний регресійний зв’язок із незалежними змінними X, для якого виконуються умови, що залишки моделі незалежні й однаково розподілені зі сталою дисперсією D та M(u) =0.

В якості функції F найпоширеніше використовуються два види функцій:

– функція логістичного розподілу:

$$F(z) = \frac{e^z}{1+e^z} \qquad (2.5)$$

за якої відповідна модель називається logit-моделлю;

– функція стандартного нормального розподілу:

$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{t^2}{2}} dt \qquad (2.6)$$

при застосуванні якої відповідна модель називається probit-моделлю.

Для знаходження оцінок параметрів β моделі використовується метод максимальної правдоподібності.

Логістичний розподіл вельми схожий на нормальний розподіл. Відповідно однозначний вибір щодо застосування одного з видів моделей є складним питанням. Відомо, що для малих вибірок та вибірок з незначним розкидом незалежних змінних висновки за logit- та probit-моделями будуть майже збігатися.

При оцінюванні якості моделі необхідно перш за все звертати увагу на значущість моделі та оцінювання параметрів.

Оцінити якість отриманої моделі можна також на основі параметра, який має назву «відношення неузгодженості» та обчислюється як відношення добутку кількості правильно класифікованих спостережень до добутку неправильно класифікованих. Відношення має бути більшим від одиниці.

Оскільки модель (2) нелінійна за параметрами, то інтерпретація цих параметрів відрізняється від лінійного випадку. Диференціюючи (2.2) по X, маємо:

$$\frac{\partial P(Bad=1)}{\partial X} = F'(X\beta')\beta = p(X\beta')\beta \tag{2.7}$$

Таким чином, граничний ефект для кожного чинника xj (j = 1, 2, ... , m) є змінною величиною і залежить від інших чинників. При використанні цієї моделі певне уявлення щодо ефекту впливу незалежних змінних можемо отримати, обчислюючи похідні (2.6) для середніх значень незалежних змінних.

Для випадку logit-моделі, коли використовується функція логістичного розподілу (4), щільність p(Xβ′) функції розподілу F(Xβ′) визначається за формулою:

$$p(X\beta') = p(z) = \frac{1}{(1+e^{z})^{2}}. \tag{2.8}$$

Отже, для визначення наближеної оцінки граничного впливу ефекту для кожного чинника xj (j = 1, 2, ... , m) необхідно скористатися формулою:

$$\Delta x_j = \frac{\beta_j}{(1+e^{\beta_0+\beta_1\bar{x}_1+\beta_2\bar{x}_2+\cdots+\beta_m\bar{x}_m})^2} \tag{2.9}$$

де Δ xj – граничний вплив ефекту для кожного чинника, xj, xj – середнє значення для кожного чинника xj (j = 1, 2, ... , m).

2.1.2 Метод максимальної правдоподібності

Існує кілька методів пошуку коефіцієнтів логістичної регресії. Оцінка максимальної вірогідності - це популярний статистичний метод, що використовується для створення статистичної моделі на основі даних та надання оцінки параметрів моделі. Метод максимальної ймовірності відповідає багатьом відомим методам статистичного оцінювання. Наприклад, припустимо, що ви зацікавлені у зростанні населення в Україні. Скажімо, у вас є дані щодо зростання деяких людей, а не всього населення. Крім того, передбачається, що зростання - це нормально розподілене значення з невідомою дисперсією та середнім значенням. Середнє значення вибірки та дисперсія приросту вибірки максимально вірогідні до середнього та дисперсії для всієї сукупності.

Для фіксованого набору даних та базової імовірнісної моделі, застосовуючи метод максимальної ймовірності, ми отримуємо значення параметрів моделі, які наближають дані до реальних. Максимальна оцінка ймовірності забезпечує унікальний та простий спосіб визначити рішення у разі нормального розподілу.

Методом максимальної правдоподібності можна оцінити невідомі параметри через максимізацію функції правдоподібності. Відповідно до цього методу вибираються такі параметри b0,...,bm рівняння регресії (2.4), що значення функції правдоподібності є максимальним на навчальній вибірці.

$$\hat{b}=\text{argmax}_b\, xL(b), \tag{2.10}$$

де $L(b)$ – функція правдоподібності; b – вектор параметрів рівняння регресії; $\hat{b}$– вектор оцінок параметрів рівняння регресії.

Випадкова величина y має розподіл Бернулі, оскільки приймає тільки два значення (0 і 1). Тоді, ймовірність настання події y = 1 рівна:

$$\begin{gathered}\Pr\{y = 1|\, x\} = f(z),\\ f(z) = \frac{1}{1 + e-z},\\ z = b0 + b1x1 + \cdots + bnxn,\end{gathered} \tag{2.11}$$

де $f(z)$ – логістична функція; xi – незалежні змінні; bi – параметри рівняння регресії.

Ймовірність настання другої можливої події y = 0 рівна:

$$\Pr\{y = 0|\, x\} = 1 - f(z). \tag{2.12}$$

Отже, функцію розподілу y при заданому x можна записати у такому вигляді:

$$\Pr\{y|\, x\} = f(b^T\, x)^y(1 - f(b^T\, x))^{1-y},\ y\ \epsilon\ \{0, 1\}. \tag{2.13}$$

Тоді, функція правдоподібності на навчальній вибірці має вигляд:

$$\mathrm{L}(b)=\prod_{i=1}^{n} Pr\{y\ =\ y(i)|\ x\ =\ x(i)\} \tag{2.14}$$

Замість максимізації функції правдоподібності можна максимізувати її логарифм:

$$\ln L(b) = \sum_{i=1}^{n} lnPr\{y\ =\ y(i)|\ x\ =\ x(i)\}\ =$$
$$=\sum_{i=1}^{n}(y(i)\ ln\ f(bTx(i))\ +(1\ -\ y(i))\ ln(1\ -\ f(bTx(i)))) \tag{2.15}$$

Наприклад, ви можете використовувати метод градієнтного спуску для максимізації цієї функції. Цей метод полягає у виконанні наступних ітерацій, починаючи з деякого початкового значення параметрів регресії b:

$$b:=b+\ \alpha^*\ \nabla\ln\ L(b) = b\ +\ \alpha\ *\ \frac{d\ \ln L(b)}{db}=$$
$$=b+\alpha*\sum_{i=1}^{n}(y(i)-f(bTx(i)))*x(i), \tag{2.16}$$

де $\alpha > 0$ – крок методу.

Також, на практиці також можливе застосування стохастичного градієнтного спуску та методу Ньютона [11].

2.2 Попередній аналіз і обробка даних для побудови скорингової моделі

2.2.1 Збір даних

Скорингова карта являє собою набір характеристик (вік, дохід, професія, досвід роботи, наявність майна тощо) позичальника та певні ваги, виражені в балах. Клієнт банку нараховує певну кількість балів залежно від інформації, яку

він повідомив про себе. Максимальна сума кредиту, яку банк готовий надати позичальнику, розраховується залежно від кількості балів.

Логістична регресія є найбільш поширеною статистичною моделлю оцінювання бінарної залежної змінної.

Оцінка коефіцієнтів логістичної регресії як балів є головним кроком у розробці карти балів. Загальна оцінка в натуральній шкалі логарифму обчислюється як сума коефіцієнтів логістичної регресії, помножена на значення незалежних змінних:

$$total\ \ score = \beta_1 x_1 + ... + \beta_k\ x_k\,, \tag{2.17}$$

де β_i – оцінки коефіцієнтів логістичної регресії; x_i – значення регресорів для і-го позичальника.

Техніка масштабування використовується для перетворення балів в лінійну шкалу. Масштабування не змінює передбачувальних властивостей карти, а лише перетворює бали в нову, просту у використанні шкалу. Оцінка за лінійною шкалою - це відношення шансів "хороших" позичальників до "поганих" [21].

Для масштабування спочатку потрібно встановити мінімальний та максимум числової шкали (наприклад, 0 і 1000). У процесі масштабування такі показники, як кількість балів, що подвоюють шанси стати "хорошим" клієнтом та значення шкали, в якій задане співвідношення шансів "погано" на "добре". Найчастіше вони використовують скоринг-карти, де кожні 20 балів подвоюють шанси стати "хорошими". Крім того, кожні 40 очок вдвічі збільшать ваші шанси стати "хорошим" позичальником. Для приведення коефіцієнтів логістичної регресії в бальні точки в лінійному масштабі використовується наступне перетворення:

$$бал = A + R * b_j, \tag{2.18}$$

де A – зміщення; R – множник.

Множник визначається по формулі:

$$R = \frac{D}{ln(2)} \ , \tag{2.19}$$

де D – бали, що подвоюють шанси.

Зміщення розраховується по формулі:

$$A = B - R * \ln(C), \tag{2.20}$$

де B – значення на шкалі балів, в якій відношення шансів складає C :1.

Отже, для побудови моделі балів необхідно зібрати достатню кількість репрезентативного зразка кредитної історії позичальників банку, або інформацію про виконання чи невиконання своїх кредитних зобов'язань. Точність прогнозу та успішність розробленої системи балів в цілому залежить від якості вихідних даних. Для побудови моделей балів необхідно використовувати достовірні та чисті дані з мінімальною кількістю записів "погано" та "добре". Кількість необхідних даних визначається вимогами статистичної значущості та випадковості, але в принципі може бути різною. Для вирішення практичних проблем розробки моделей балів, експерти з банківських рахунків рекомендують використовувати щонайменше 2000 «поганих» та 2000 «хороших» записів клієнтів, які випадковим чином вибираються із загальної історії відповідного банку або кредитних бюро. Крім того, додатково можуть знадобитися 2000 запитів на бал за допомогою спеціальних методів оцінювання для аналізу причин відхилення. Грубі дані для побудови моделей балів можуть включати внутрішні дані з банківських профілів позичальників та зовнішньої кредитної історії.

2.2.2 Визначення та обробка пропусків

Зазвичай історичні дані характеризуються відсутністю якихось необхідних значень або, навпаки, наявністю невірних значень і не можуть описати ту чи іншу характеристику. Це можуть бути поля, значення яких більше не використовуються, або не були зафіксовані, або, можливо, не були доступні або не були заповнені позичальниками, тобто пропущені значення; а також неправильно введені дані, викиди або значення, які є дуже помітними, тобто помилковими, неправильними даними. Існує кілька методів обробки даних із такими значеннями, наприклад:

а) виключити з аналізу всі дані з відсутніми значеннями, оскільки аналіз проводиться за всіма змінними. У разі роботи з реальними фінансовими даними цей метод в більшості випадків видаляє занадто багато даних;

б) виключити характеристики або записи, у яких частка відсутніх значень перевищує певний поріг (наприклад, більше 20%);

в) включити в аналіз нову характеристику (ідентифікатор), яка відображає наявність пропуску на атрибут клієнта;

г) замінити пропущені значення на основі середнього або прогнозування (наприклад, дерева рішень або методи регресії) або статистичних спеціалізованих методів (синтетичний розподіл).

2.2.3 Визначення цільової та незалежних змінних моделі

Мета побудови скорингової моделі залежить від вибору цільової (залежної) змінної. Цілі можуть бути різними: загальні - зменшення збитків за щойно виданими позиками, конкретні - зменшення непогашених позик за

наданими позиками протягом 4 місяців після прийняття рішення про випуск. Залежна змінна може приймати кількісні та якісні значення. Найчастіше залежна змінна має категоричний тип вимірювання і приймає дві категорії: "хороший" та "поганий" клієнт. До категорії «поганих» зазвичай відносять клієнтів із заборгованістю 90 днів і більше.

Незалежні змінні в побудові моделі кредитного балотування - це дані заявки на позику, такі як: соціально-демографічні дані клієнта (вік, стать, сімейний стан, наявність дітей, посада, дохід тощо), інформація про позику (термін повернення позики, сума позики, сума першого внеску, мета кредиту тощо). Також під час подання заявки клієнт використовує дані бюро кредитів як основне джерело даних для формування незалежних змінних. Основні особливості, які може надати Бюро, це: рейтинг клієнтів, детальна інформація про наявність позик в інших банках, інформація про минулі або повністю погашені позики в минулому, наявність інших банківських послуг та продуктів від клієнта. Також для формування незалежних змінних використовується внутрішня кредитна історія позичальника: залишок поточного рахунку, поточний борг, кількість рахунків, кількість попередніх позик банку, максимальна сума боргу на попередніх кредитних рахунках.

Тому незалежні вхідні змінні досить різноманітні і можуть бути представлені в різних одиницях залежно від здатності об'єктивно вимірювати вибрані характеристики. Для вирішення практичних завдань моделі оцінки можуть бути побудовані з таких типів незалежних змінних: лише з категоріальною, тільки з кількісною, одночасно з кількісною і категоріальною змінними. Найбільш використовувані змінні в побудові моделей оцінки - це категоріальні змінні. Основними перевагами категоризації кількісних змінних при побудові бальних моделей є: полегшення переробки викиді та екстремальних значень кількісних змінних, показ складних нелінійних зв’язків [20].

2.2.4 Відбір найбільш вагомих змінних

Кінцева модель повинна включати лише найбільш значущі незалежні змінні, які, будуючи скорингову модель, матимуть найбільш прогнозні характеристики.

Для оцінки прогнозної сили атрибутів використовують зважену сукупність - WOE (Вага доказів). WOE вимірює статистичну значимість кожного класу змінної і обчислюється як:

$$\text{WOE}= \ln\left(\frac{d_i^{(1)}}{d_i^{(2)}}\right), \tag{2.16}$$

де $d_i^{(1)}$-відносна частка «хороших» кредитів в і -й категорії; $d_i^{(2)}$ – відносна частка «поганих» кредитів в і-й категорії.

При розрахунку прогнозуючої сили характеристики в цілому значення WOE для кожного атрибута агрегуються в показник інформаційного значення, або індекс IV (Information Value).

Показник інформаційного значення прийнято використовувати в кредитному скорингу для оцінювання ступеня взаємозв'язку між залежною змінною і незалежними, IV обчислюється за формулою:

$$\text{IV}=\sum_{i=1}^{k} d_i^{(1)} - d_i^{(2)} * \text{WOE}, \tag{2.17}$$

де $d_i^{(1)}$- відносна частка «хороших» кредитів в і-й категорії; $d_i^{(2)}$ – відносна частка «поганих» кредитів в і-й категорії; WOE – значення зваженої сукупності і-ї категорії; k – кількість категорій незалежної змінної.

Корисність змінної при побудові скорингової моделі визначається її інформаційним значенням, чим вище змінна, тим вона корисніша. Також

доцільно враховувати наступні правила при виборі змінних для побудови скорингової моделі (табл. 2.2).

Таблиця 2.2 — Оцінка значущості незалежної змінної за значенням IV

Значення IV	Прогнозна здатність
менше 0.02	Не має
від 0.02 до 0.1	Низька
від 0.1 до 0.3	Середня
від 0.3 до 0.5	Добра
більше 0.5	Дуже добра

2.2.5 Створення навчальної та тестової вибірки

Важливим кроком у створенні моделі оцінки є затвердження її реальними даними та перевірка. Ступінь валідації моделі вказується на її здатність правильно класифікувати об’єкти, здатність моделі відрізняти "хороших" позичальників від "поганих". Модель повинна правильно передбачати не тільки навчальний зразок, але і на практиці при його застосуванні. Найпоширеніша стратегія валідації моделі адекватності полягає у формуванні двох зразків навмання: на ній будується навчальна модель, а тестова модель призначена для тестування моделі. Перевірка моделі зазвичай проводиться з використанням навчальних та тестових зразків у пропорціях приблизно 75-85% та 25-15% відповідно від вихідних даних. Якісна модель повинна демонструвати прийнятну здатність до прогнозування як у навчанні, так і в тестових зразках. Подібні статистичні дані, розраховані на навчальних та тестових зразках, є свідченням того, що модель буде більш стабільною на практиці та дасть адекватні прогнози.

Більш складна стратегія валідації моделі передбачає формування трьох або більше зразків: перша вибірка використовується для оцінки параметрів моделі, друга вибірка використовується для тестування моделі, якщо отримані значні відхилення результатів від навчальних та тестових зразків, вони видаляють викиди або змінні, що впливають на відхилення, і будується нова модель для об'єднання першого та другого зразків, а результати нової моделі тестуються на третій вибірці.

2.3 Методи та підходи щодо оцінювання скорингових моделей

2.3.1 Прості методи оцінки параметрів моделі

Набір статистичних критеріїв, інструментів та процедур використовується для перевірки та оцінки якості моделей. Однією з найпоширеніших оцінок якості моделі в завданнях прогнозування є середня абсолютна помилка (MAE) та середня квадратична помилка (MSE):

$$\text{MSE}=\frac{1}{N}\sum_{i=1}^{N}(d_i - y_i)^2,$$

$$\text{MSE}=\frac{1}{N}\sum_{i=1}^{N}|\text{d}_\text{i} - \text{y}_\text{i}|^2, \qquad (1.28)$$

де N – кількість спостережень; d_i – реальне значення цільової змінної і-го спостереження; y_i – прогнозне значення.

Середньоквадратична помилка набагато більш чутлива до великих відхилень, ніж середня похибка, і тому більш чутлива до викидів. Використовуючи будь-яку з двох помилок, буде корисно проаналізувати об'єкти, які дають найбільшу помилку. Стандартна помилка дає хороші результати порівнюючи дві моделі або контролю якості на етапі навчання, але

не дозволяє робити висновки про адекватність цієї моделі. Наприклад, значення MSE = 10 є дуже поганою характеристикою моделі, цільова змінна якої приймає значення від 0 до 1, і навпаки дуже хорошим, якщо цільова змінна знаходиться в інтервалі (10000, 100000). У таких випадках замість середньоквадратичної помилки використовується коефіцієнт детермінації, або коефіцієнт :

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(d_i - y_i)^2}{\sum_{i=1}^{N}(\bar{y} - y_i)^2}, \tag{2.19}$$

де $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ – середнє значення цільової змінної.

Якщо коефіцієнт детермінації ближчий до одиниці, то модель є адекватною і має хороші прогнозні якості, але якщо наближається до нуля, прогностичну здатність такої моделі можна порівняти з якістю з постійним прогнозуванням.Загальна точність моделі (CA – англ. Common Accuracy) – обчислюється як відношення вірно спрогнозованих значень до загальної кількості значень N:

$$\text{CA} = \frac{\text{к-сть вірно спрогнозованих значень}}{\text{N}}, \tag{2.20}$$

В ідеалі CA повинен прямувати до 1.

2.3.2 ROC-крива та індекс Gini

Ефективним способом оцінки точності моделі, що класифікує вхідні дані на два класи є побудова і аналіз ROC-кривої (Receiver perating Characteristic. ROC-крива відображає залежність долі правильно класифікованих позитивних прикладів від долі неправильно класифікованих негативних прикладів. Перші

частки називаються істинно позитивними, а інші частинки - неправильно негативними. Також передбачається, що в класифікатора є певний змінний параметр, зміна якого дозволить отримувати ту чи іншу раз биття. Цей параметр називається поріг відсікання (cut-off value). в залежності від його значення в результаті будуть різні значення помилок I і II роду.

Розглянемо більш детально таблицю спряженості (confusion matrix), яка будується на основі класифікації моделі і фактичної приналежності прикладів класам (табл.2.3), де:

а) TP (True Positives) – вірно класифіковані позичальники, що повернули кредит;

б) TN (True Negatives) – вірно класифіковані позичальники, що не повернули кредит;

в) FN (False Negatives) – позичальники, що повернули кредит, класифіковані як ті що не повернули (помилка I роду);

г) FP (False Positives) – позичальники, що не повернули кредит, класифіковані як ті що повернули (помилка II роду).

Таблиця 2.3 – Таблиця спряженості

	Дійсна приналежність	
Результат прогнозування	Негативний	Позитивний
Негативний	TN (Істинно негативний)	FN (Хибно негативний)
Позитивний	FP (Хибно позитивний)	TP (Істинно позитивний)

Доля істинно позитивних прикладів (True Positives Rate) визначається за формулою:

$$TPR = \frac{TP}{TP+FN}*100\%, \quad (2.21)$$

Доля хибно позитивних прикладів (False Positives Rate):

$$FPR = \frac{FP}{TN+FP}*100\%, \quad (2.22)$$

Для побудови ROC-кривої вводиться ще наступні поняття: чутливість і специфічність моделі. За допомогою цих понять визначається об’єктивна значущість будь-якого бінарного класифікатора.

Чутливість (Sensitivity) моделі – це доля істинно позитивних випадків:

$$Specificity = \frac{TN}{TN+FP} * 100\% = \ 100\% - FPR, \qquad (2.23)$$

Зазвичай найчастіше правильно класифікує позитивні приклади модель з високою чутливістю, а модель з високою специфічністю навпаки, краще справляється з виявленням негативних прикладів. ROC-криву (рис. 2.4) отримують у наступний спосіб:

1) спочатку розраховуємо значення чутливості та специфічності длякожного значення порога відсікання, змінюючи його від 0 до 1 з певним кроком dx (наприклад, 0.01);
2) будуємо графік залежності: по осі ординат відкладається значення чутливості, по осі абсцис відкладається значення розраховане наступним чином: 100% - Specificity.

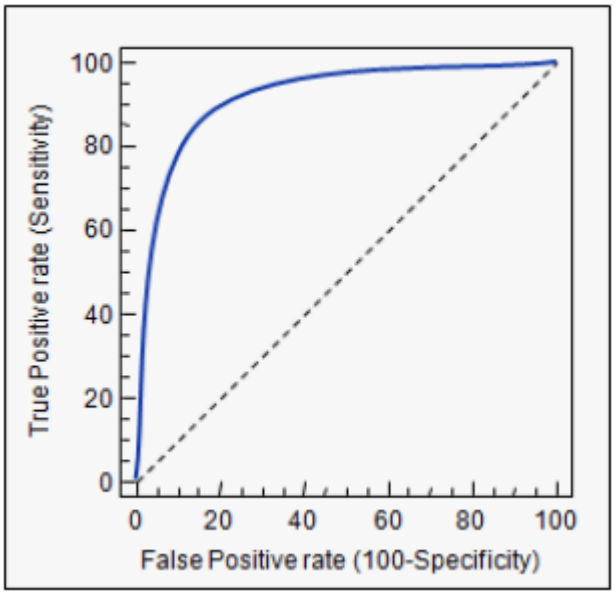

Рисунок. 2.4 — Приклад побудови ROC-кривої

Вибір оптимального значення порогового значення залежить від того, яка помилка є більш допустимою, першого чи другого роду при класифікації. При зниженні порога в моделі буде переважати чутливість, тобто здатність моделі правильно виявляти позичальників, що будуть мати прострочені платежі. Також в якості оптимального порогу відсікання можна обрати точку балансу між чутливістю і специфічністю.

Для порівняння різних моделей (або моделей з різними параметрами) використовується площа під ROC-кривою – AUC (Area Under Curve). Площа AUC змінюється в діапазоні від 0.5 до 1 (табл. 2.4).

Таблиця 2.4 — Оцінка якості моделі за значенням площі AUC

Значення AUC	Якість моделі
0.9-1	Відмінна
0.8-0.9	Дуже добра
0.7-0.8	Добра
0.6-0.7	Середня
0.5-0.6	Незадовільна

Слід зазначити, що призначення показника площі під кривою лише для порівняльного аналізу моделей між собою. Показник площі під кривою не несе ніякої інформативності про чутливість і специфічність моделі.

При аналізі якості моделі з використання значення площі під ROC-кривою зазвичай розраховують індекс Джині. Цей показник трансформує значення площі під кривою в діапазон значень від 0 до 1, чим вище його значення, тим вище дискримінуюча здатність моделі. Індекс Джині розраховується наступним чином:

$$GINI= 2*AUC –1, \quad (2.24)$$

де AUC – площа по ROC-кривої [22].

2.3.3 Статистика Колмогорова-Смірнова

Для оцінювання прогнозної здатності моделі в кредитному скорингу використовується тест або статистика Колмогорова-Смірнова. У цьому тесті проводиться перевірка статистичної гіпотези, що дві довільні вибірки є складовою однієї генеральної сукупності. У випадку скорингу відбувається порівняння двох кумулятивних розподілів скорингових балів «хороших» і «поганих» позичальників. Значення статистики Колмогорова-Смірнова розраховується як максимальна різниця між значеннями кумулятивних функцій розподілу «поганих» і «хороших» клієнтів:

$$KS = \max_x |Fm(x) - Gn(x)| \ *100\% \quad (2.25)$$

де F (m) x – кумулятивний розподіл скорингового балу для «поганих» клієнтів;клієнтів; Gn (x) – кумулятивний розподіл скорингового балу для «хороших» m – кількість «поганих» клієнтів; n– кількість «хороших» клієнтів.

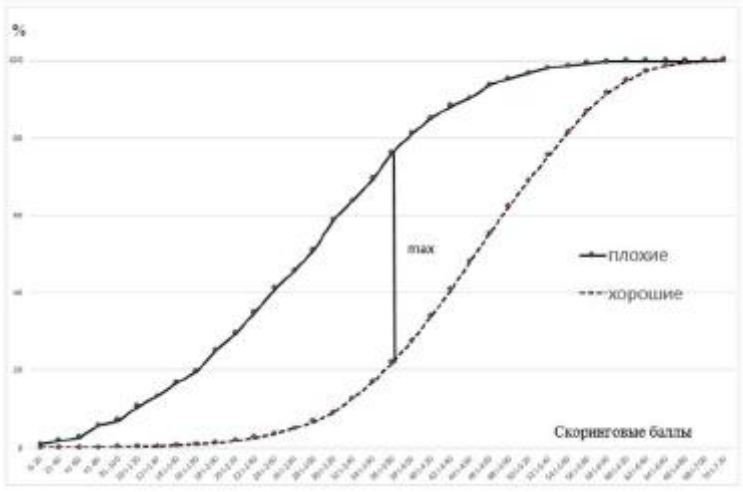

Рисунок. 2.5 — Ілюстрація розрахунку статистики Колмогорова-Смірнова

Визначимо алгоритм обрахунку статистики Колмогорова-Смірнова і перевірки гіпотези рівності двох функцій розподілу. Ранжуємо клієнтів в порядку збільшення скорингового балу і групуємо їх. Ознакою групування

виступає отриманий скоринговий бал. Далі в кожній отриманій групі клієнтів розраховуємо наступні показники:

1. кількість «хороших» клієнтів;
2. кількість «поганих» клієнтів;
3. відношення шансів «поганих» до «хорошим» клієнтів;
4. відсоток «поганих» та «хороших» позик;
5. кумулятивну суму «поганих» та «хороших» позик;
6. кумулятивний відсоток «поганих» та «хороших» позик;
7. загальний кумулятивний відсоток поганих позик від їх загальної кількості;
8. різницю між кумулятивними відсоткам поганих і хороших позик.

Після чого необхідно вирахувати максимальну різницю між кумулятивним відсотком «хороших» і «поганих» позик і розрахувати за формулою значення статистики Колмогорова-Смірнова. Обчислене значення порівнюється із значенням взятим з таблиці розподілу Колмогорова- Смірнова з вибраним рівнем значущості або при числі «поганих» і «хороших» клієнтів більше 80 можна обрати наближене порогове значення, що розраховується за наступною формулою:

$$z(a)=\sqrt{((m+ n)/mn)}, \qquad (2.26)$$

де z(a) – значення, що відповідає обраному рівнем значущості.

Якщо розраховане значення статистики за формулою (2.25) менше значення по таблиці або значення розрахованого за формулою (2.26), то гіпотеза рівності двох функцій розподілів відкидається.

Значення статистики Колмогорова-Смірнова можуть змінюватися в діапазоні від 0 до 100. Високі значення статистики Колмогорова-Смірнова, говорять про кращу здатність моделі до класифікації. Зазвичай значення

статистики Колмогорова-Смірнова, лежать в діапазоні від 20-25 до 75-80, а крайні значення статистика не приймає [21].

Висновки до розділу 2

В другому розділі було проведено огляд існуючих математичних методів прогнозування, які можна використовувати для оцінювання кредитоспроможності фізичних осіб, а саме лінійна імовірнісна модель, логістична регресія та скорингова карта. Також визначено алгоритми, які використовують для побудови та навчання логістичної регресії.

Досліджено процес попереднього аналізу і обробки даних для побудови скорингової моделі, який включає в себе: збір даних, обробку пропусків, визначення цільової змінної, відбір найбільш значущих змінних, формування навчальної та тестової вибірки.

Також були розглянуті критерії оцінки якості отриманих прогнозуючих моделей: загальна точність моделі, ROC-крива, індекс Gini, а також помилки 1-го й 2-го роду. Останні особливо важливі з погляду фінансової установи, оскільки помилка 1-го роду означає, що банк втратить певну суму грошей, а помилка 2-го роду означає, що банк не доодержить деякий прибуток.

3 СИСТЕМА ПІДТРИМКИ ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ ВИЗНАЧЕННЯ КРЕДИТОСПРОМОЖНОСТІ ФІЗИЧНИХ ОСІБ

В якості практичного застосування СППР розроблено комп’ютерну програму «Logist analytics», опису якої присвячений даний розділ. Дана програма призначена для прогнозування кредитоспроможності позичальників спираючись на методи регресійного моделювання. Продукт створено мовою програмування C# в середовищі розробки Microsoft Visual Studio 2012.

Програмний продукт дозволяє користувачам незалежно від рівня підготовки проводити необхідний збір даних для побудови прогнозуючої моделі, розробляти скорингову модель, та отримувати у результаті статистичні характеристики й прогнозні дані.

Інтерфейс користувача інтуїтивно зрозумілий і створений таким чином, щоб провести оператора від моменту завантаження вхідних даних до виведення та збереження результатів [29].

Для обчислення коефіцієнтів рівняння логістичної регресії в програмі застосовано метод градієнтного спуску. Спеціально для полегшення реалізації алгоритму на основі методу максимальної правдоподібності був розроблений модуль Matrix, який представляє собою набір процедур для роботи з матрицями.

За технічним рівнем Logist analytics належить до настільних програмних продуктів, система не розрахована на мережеву роботу.

3.1 Аналіз архітектури системи

На рисунку 3.1 наведена архітектура розробленої СППР. Як можна побачити з цієї структури СППР представляє собою широкий комплекс засобів для аналізу та обробки даних.

Рисунок 3.1 — Структура системи підтримки прийняття рішень

Через пристрої вводу-виводу аналітик має змогу завантажувати вхідні дані в розроблену систему підтримки прийняття рішення. Тому підсистема вводу-виводу функціонально пов’язана з підсистемою інтерфейсу користувача.

Підсистема інтерфейсу користувача дає можливість здійснення зв’язку між користувачами СППР та внутрішніми підсистемними одиницями, що дає змогу на ввід та вивід інформації для ОПР і експертів, а також надає доступ до зовнішніх запам’ятовуючих пристроїв ПЕОМ. Інтерфейс дозволяє операторові вводити інформацію, дані, команди, запити в систему та одержувати вихідну інформацію в зручному для сприйняття вигляді.

Підсистема зберігання інформації складається з бази даних, яка призначена для накопичення статистичних даних для їх подальшого опрацювання та використання.

Підсистема обробки інформації призначена для перевірки даних, що надходять, на наявність пропусків, заповнення цих пропусків, категоризації безперервних даних.

Блок аналізу та прогнозування складається з трьох підсистем: підсистема побудови моделі, підсистема прогнозування, підсистема оцінки якості моделі.

Підсистема побудови моделі реалізована за допомогою логістичної регресії з використанням методу максимальної правдоподібності.

Підсистема оцінювання якості моделі підраховує статистичні показники розробленої моделі, які характеризують прогнозуючу здатність моделі.

Підсистема виведення результатів містить в собі набір графіків, таблиць та звітів для прийняття рішення експертом. Представлення результатів прогнозування та критеріїв оцінювання моделі дає можливість оцінити доцільність використання отриманої моделі для прогнозування.

У розробці системи підтримкм прийняття рішення використано технології .Net та середовище Microsoft Visual Studio 2012.

При застосуванні методу максимальної правдоподібності було розроблено модуль Matrix з набором процедур для роботи з матрицями.

На сьогодні більшість людей використовує операційні системи класу Windows, через що інтерфейс користувача був створений на базі технологій WinForms з вводом вікон та іконок. Що у результаті спростило роботу споживача, адже стало непотрібним вивчення великої кількості команд для виконання процедур аналітики, а отримані результати одразу ж освітлені у зручному вигляді.

3.2 Інструкція з експлуатації програмного продукту

Для роботи програмного продукту необхідна наявність персонального комп’ютера з наступними мінімальними характеристиками:

- операційна система Windows 7/8/10;
- тактова частота процесора 1 ГГц;
- оперативна пам’ять розміром 512 Мбайт;
- вільний дисковий простір: 5 Мбайт для розміщення виконавчого файлу, вхідних даних і результатів роботи;
- монітор з розподільчою здатністю 1024×768;
- інсталяція .Net Framework версії 4.5.

Програмний продукт, який реалізовано в межах магістерської роботи, розроблено для роботи в операційній системі MS Windows. Усі помилкові введення даних опрацьовуються системою та інформують про це споживача повідомленнями.

Основний робочий екран программного продукту має структуру зображену нижче на рисунку 3.2 та складається з трьох елементів:

- меню користувача («Область 1» на рис. 3.2), що містить команди для завантаження файлів даних, опрацювання даних, розробки моделі прогнозування та збереження файлів;
- дерево проекта («Область 2» на рис. 3.2) має три батьківські вершини, які описують процеси роботи системи: Data files (Файли даних), Prediction models (Прогнозуючі моделі), Results (Результати);
- область робочої програми («Область 3» на рис. 3.2), в якій відкриваються вікна з дерева проекту.

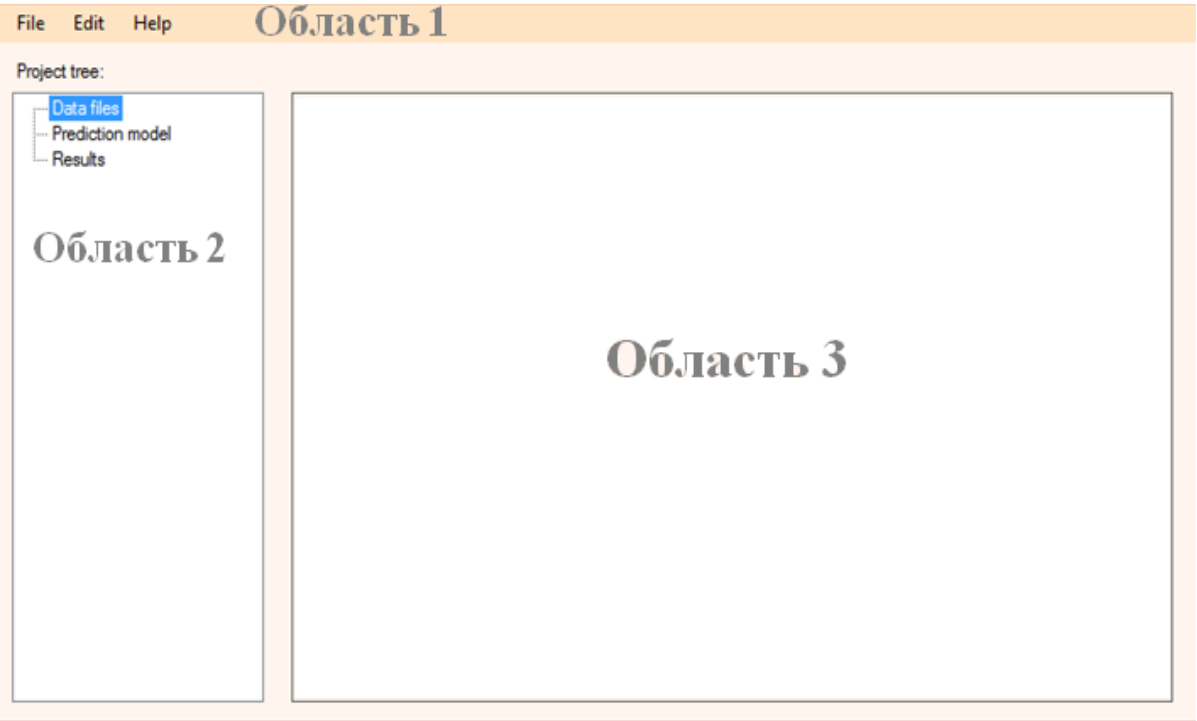

Рисунок 3.2 — Структура головного робочого екрану

3.2.1 Завантаження даних

Програма може обробити будь-яку кількість наборів даних в форматах – xls/xlsx/csv. Для підготовки даних до аналізу рекомендується використати програму Microsoft Excel. Статистичні дані сприймаються як числовими значеннями, так і строковими. На головному меню переходимо до вікна завантаження файла: File (Файл) → Open Data file (Відкрити файл даних). Після відкриття стандартного діалогово вікна Windows вибору директорії з файлами вказуємо формат файлу з даними та безпосередньо сам файл даних.

Як видно з наведеного далі рисунка (рис 3.3), вікно з попереднім оглядом метаданих про необхідний файл містить формат файлу та шлях його розташування в пам’яті комп’ютера. Наступним кроком є відзначити чи містить набір даних перший рядок з назвами змінних, встановивши або забравши відповідний прапорець.

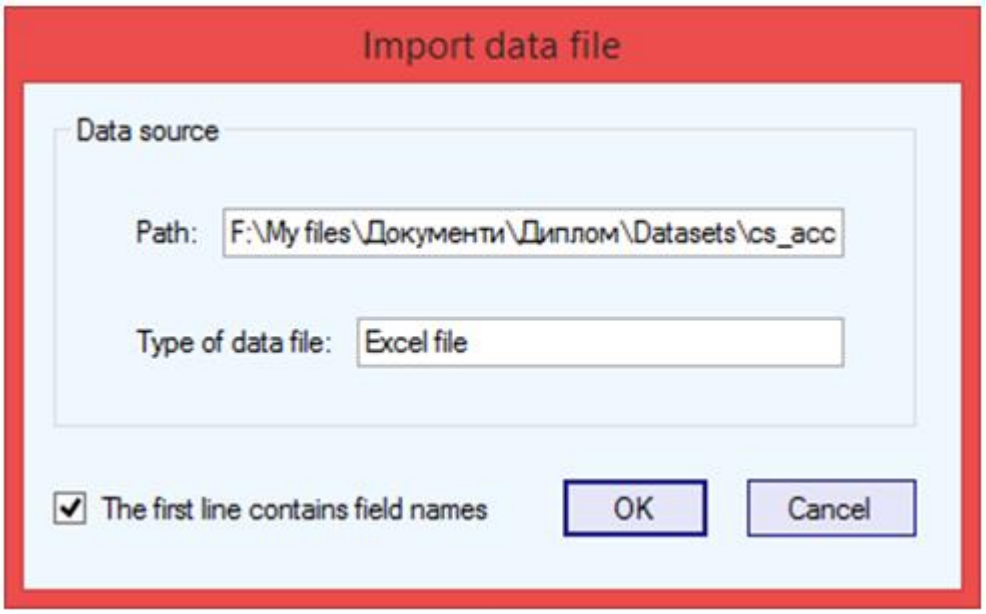

Рисунок 3.3 — Вікно з інформацією про обраний файл

Натиснувши кнопку «ОК» назва файлу добавляється до дерева проекту в вершину Data files (Файли даних), дані вставляються до аналітичної системи і відтворюються на головному екрані програми, як видно на нижченаведеному скрині (рис.3.4).

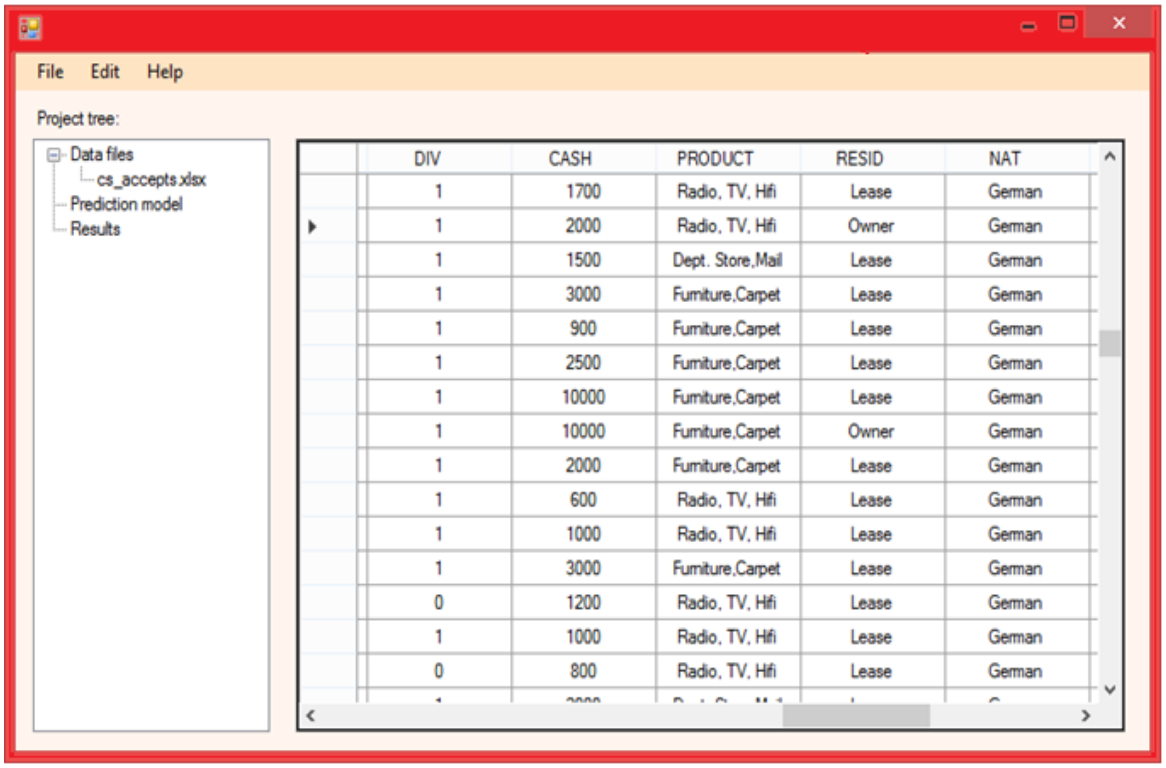

Рисунок 3.4 — Завантажені в програму дані

3.2.2 Опрацювування вхідних даних

Розроблена програма дозволяє споживачам виконувати підготовчий етап обробки даних, що складається із заповнення пропусків та дискретизація неперервних значень.

З першу відкриємо вікно для заповнення пропусків даних, для цього на головному меню обираємо Edit (Редагувати) → Missing value (Відсутні значення).

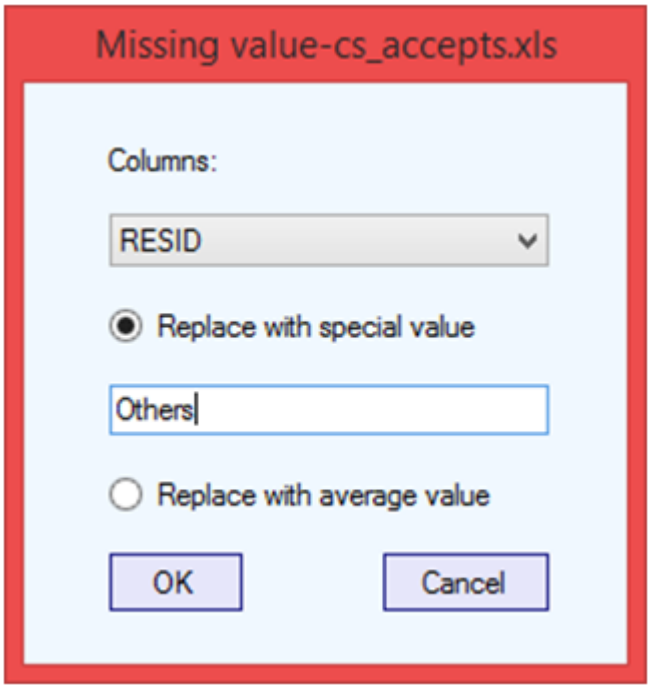

Рисунок 3.5 — Вікно опрацювання пропусків даних

У наступному вікні треба вказати стовпчик з пропусками, а також метод заповнення пропусків: спеціальним значенням, відповідно обравши його, або середнім значенням (для неперервних змінних) (рис.3.5). З натисканням кнопки «ОК» та здійснення процесу, з’явиться інформаційне вікно (рис. 3.6), в якому буде вказана кількість порожніх значень в обраному стовпчику.

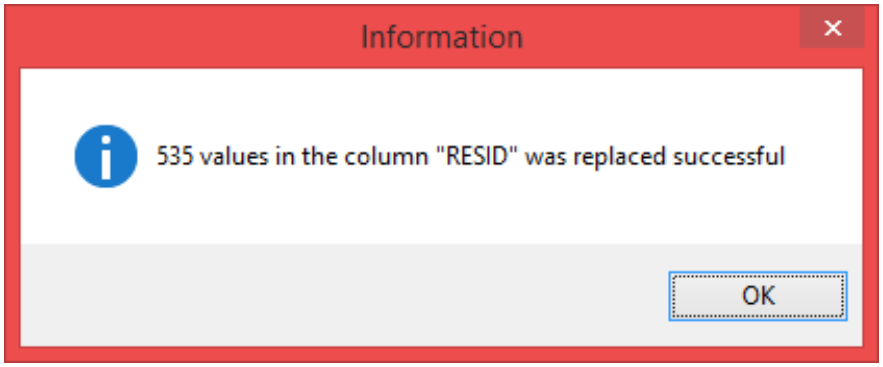

Рисунок 3.6 — Інформаційне вікно після заповнення пропусків

Щоб відкрити вікно, яке дозволяє трансформувати неперервні дані в категоріальні, слід на головному меню обрати Edit (Редагувати) → Discretization (Дискретизація). Далі відкривається вікно на дискретизацію даних, яке наведено нижче (рис. 3.7).

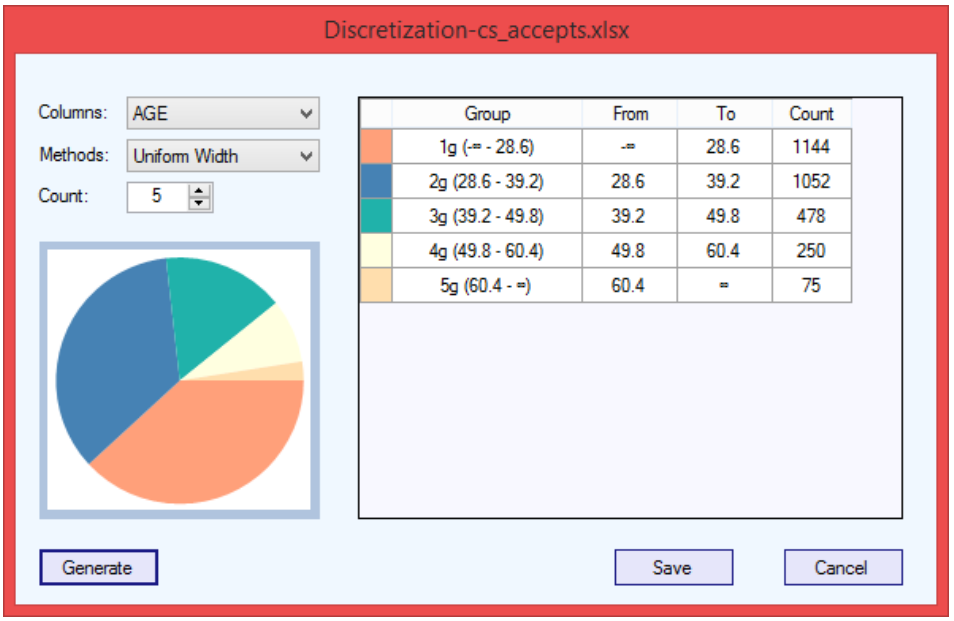

Рисунок 3.7 — Вікно для дискретизації даних

У данному вікні вказуємо неперервну змінну, яку слід дискретизувати, кількість категорій (від 2 до 20), та один з методів дискретизації: одна ширина проміжків (Uniform Width); одна кількість спостережень (Uniform Count).

Далі натиснемо кнопку «Generate» (Згенерувати), і у вікні відтвориться кругова діаграма розподілу спостережень за категоріями та таблиця з даними по категоріям, що видно з наведених нижче рисунків (рис. 3.8 – 3.9).

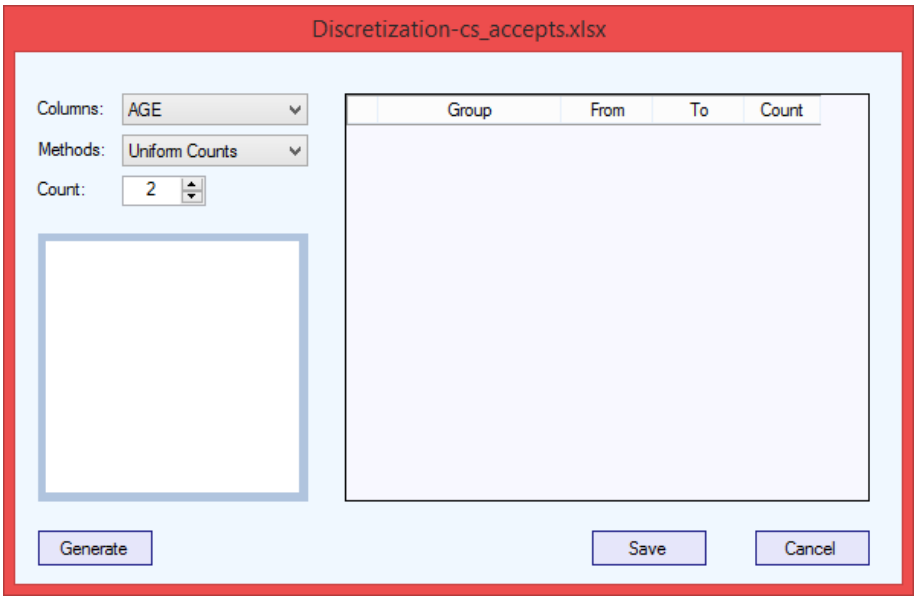

Рисунок 3.8 —Дискретизація даних методом Uniform Width

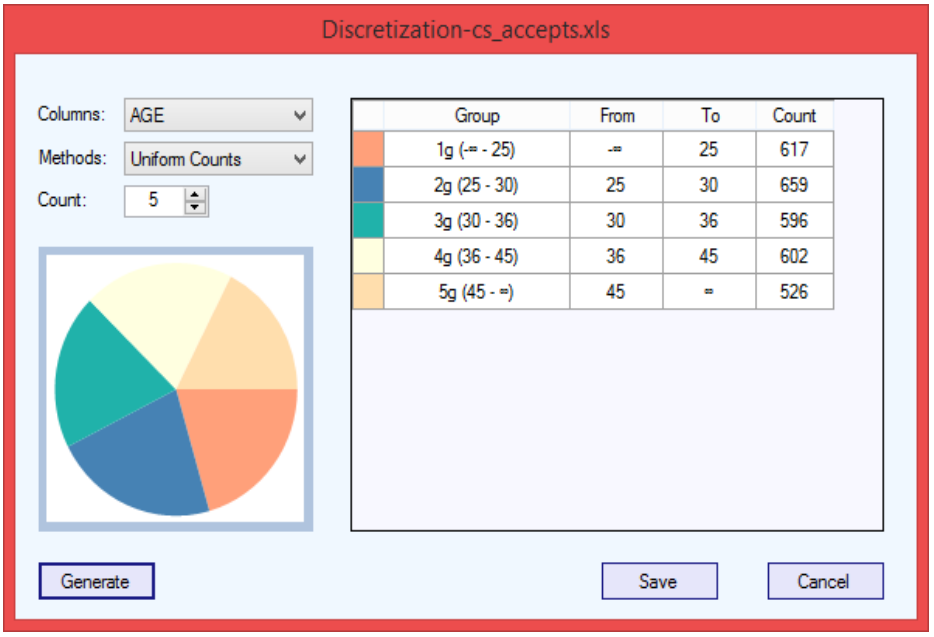

Рисунок 3.9 —Дискретизація даних методом Uniform Count

Для збереження даних введено кнопку «Save» (Зберегти), після натиснення на яку відкриється головне вікно програми, де обраний стовпчик для дискретизації з неперервними даними заміниться на відповідний йому стовпчик зі значеннями категорій.

3.2.3 Розробка прогнозованої моделі

Головним завданням програми є розробка прогнозуючої моделі на основі логістичної регресії. Дана модель будується у два кроки. З першу на головному меню обираємо Edit (Редагувати) → Build model (Побудова моделі), тим самим репрезентуючи вікно для побудови (рис.3.10).

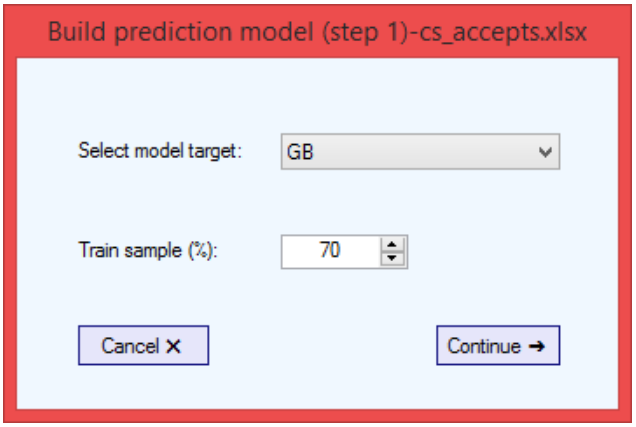

Рисунок 3.10 — Вікно побудови моделі

У вищезазначеному вікні обираємо цільову змінну прогнозування, яка є бінарною, і встановивши відсоток даних, який буде опрацьований для навчання моделі, натискаємо кнопку «Continue» (Продовжити), і переходимо на завершаючий крок розробки моделі. Якщо обрана цільова змінна виявиться не бінарною програма видасть помилку (рис. 3.11).

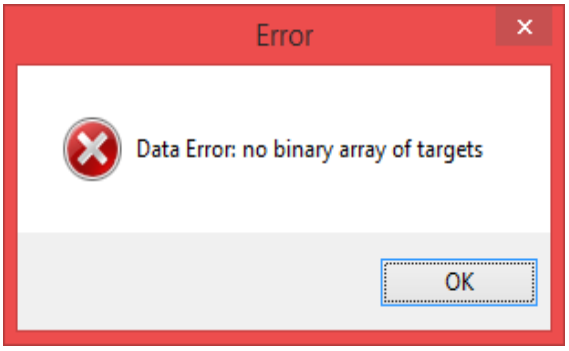

Рисунок 3.11 — Повідомлення про невідповідність цільової змінної

Якщо цільова змінна приймає коректні значення відкриється наступне вікно побудови моделі (рис. 3.12).

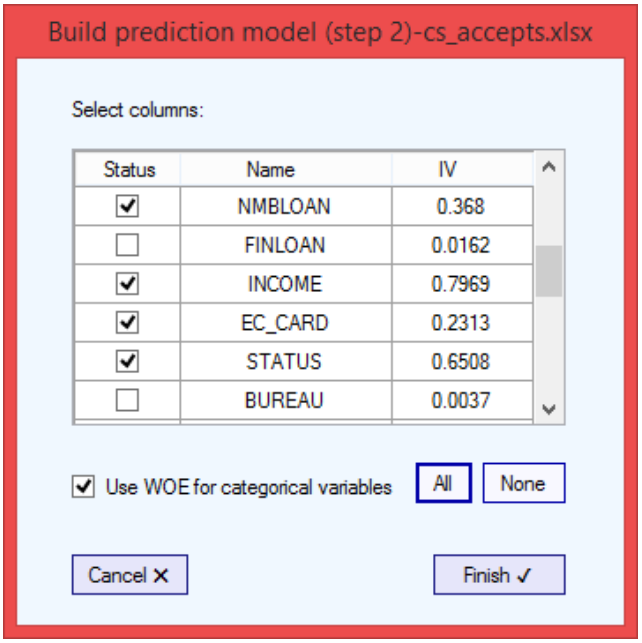

Рисунок 3.12 — Вікно побудови моделі

У вищезазначеному вікні міститься набір стовпчиків даних з їх інформаційним значеням. Програма автоматично виділяє прапорцями стовпці з інформаційним значенням вище за 0,03. Поряд з тим, споживач в змозі сам обрати необхідні стовпці для побудови моделі, а також виділити всі змінні обравши «All» (Все), або забрати усі прапорці всіх змінних натиснувши «None» (Нічого).

Логістична регресія містить лише числові значення, тому здійснено перекодування категоріальних змінних в числові за допомогою порядкової нумерації категорій, або з використанням значень коефіцієнта зваженої сукупності (WOE). За замовчуванням діє перший варіант кодування, для другого варіанту слід вказати прапорець «Use WOE for categorical variables» (Використання WOE для категоріальних змінних).

Після налаштування всіх параметрів необхідно натиснути «Finish» для побудови моделі та висвітлення результатів прогнозування.

3.2.4 Виведення результатів прогнозування

Результатом побудови моделі є два вікна, які переходять в дерево проекту у вузли «Prediction model» (Прогнозуюча модель) та Results (Результати).

Початкове вікно під назвою «Model» (Модель) містить таблицю з розробленою логістичною моделлю та оцінками параметрів логістичної регресії (рис. 3.13). Крім цього, у вікні показані обчислені статистичні коефіцієнти оцінки якості моделі, такі як: загальна точність моделі (CA), індекс GINI і графічна характеристика – ROC-крива.

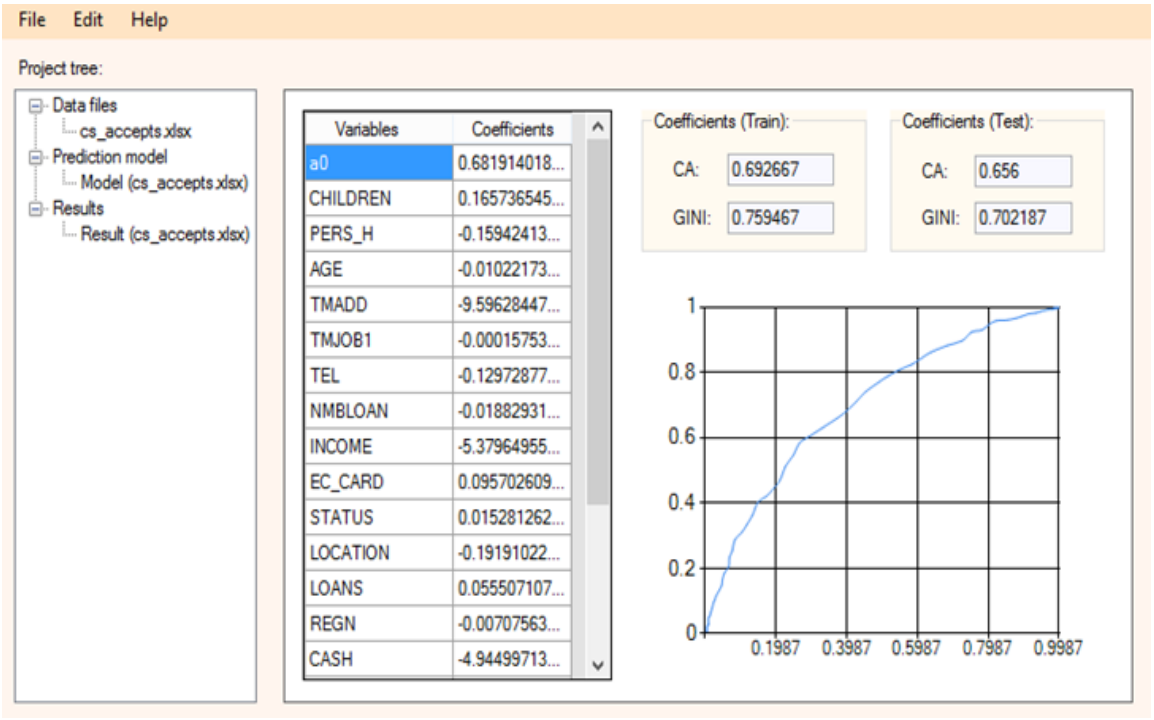

Рисунок 3.13 — Вікно з результатами побудови моделі

Наступне вікно має назву «Result» (Результат) та показує таблицю з даними, що входили до набору прогнозу, опрацьовані значення побудованої моделі, та інформацію про наявність похибок першого й другого роду (рис. 3.14).

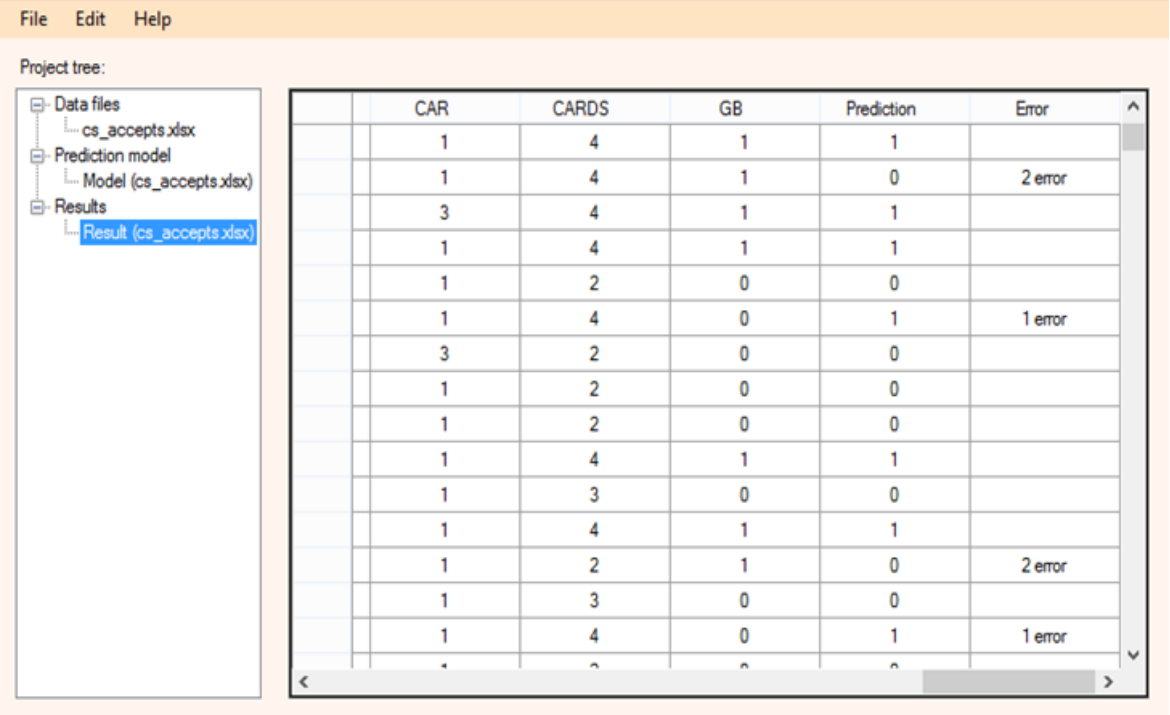

Рисунок 3.14 — Вікно з результатами прогнозування

За потреби програма може зберегти одержані результати прогнозування та параметри моделі. Відкривши на головній сторінці програми відповідне вікно з інформацією, яку необхідно зберегти, натискаємо File (Файл) → Save (Зберегти), після чого у діалоговому вікні вказуємо ім’я файлу та шлях місця розташування для збереження.

3.3 Результати апробації програмного продукту

У межах магістерської роботи, був використаний набір даних по клієнтах з трьома тисячами записів, по яким закінчився строк кредитування. Обраний набір даних містить у собі інформацію по сімнадцяти показникам (з анкетного заповнення) на одного клієнта. Опишемо кожен показник більш детально (табл. 3.1).

Таблиця 3.1 – Опис змінних набору даних

Назва змінної	Опис
AGE	Вік
CAR	Наявність транспортного засобу
CARDS	Тип банківської карти
CASH	Запитувані грошові кошти
CHILDREN	Кількість дітей
ECARD	Наявність банківської карти
FINLOAN	Кількість закритих кредитів
GB	«Хороший» чи «поганий» клієнт
INCOME	Дохід особи
LOANS	Кількість відкритих кредитів
NAT	Національність особи
NMBLOAN	Кількість кредитів, наданих цим банком
PRODUCT	Ціль кредиту
REGN	Регіон
RESID	Тип місця проживання
STATUS	Сімейний стан
TITLE	Стать особи

Вихідний набір даних було розбито на навчальну та тестову вибірки розміром 70% та 30% відносно даного набору.

Для аналізу запропонованого методу трансформації категоріальних змінних у числові було побудовано дві скорингові моделі: з використанням порядкової нумерації категорій, та з використанням коефіцієнтів WOE (рис. 3.15-3.16).

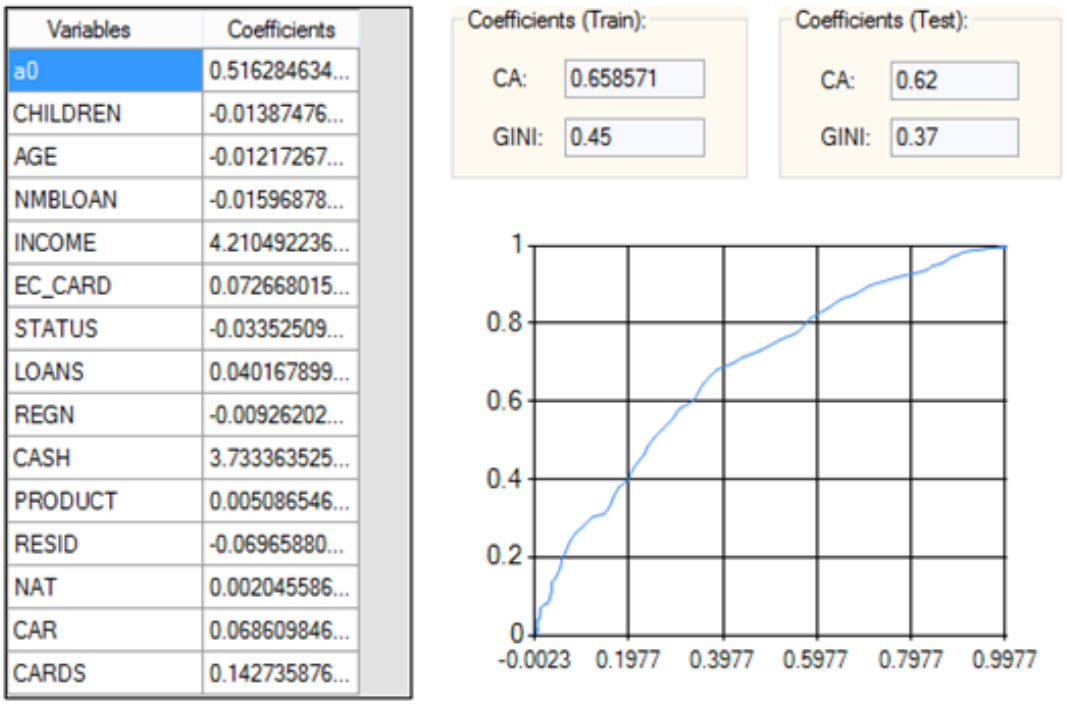

Рисунок 3.15 — Результати прогнозування без використання WOE

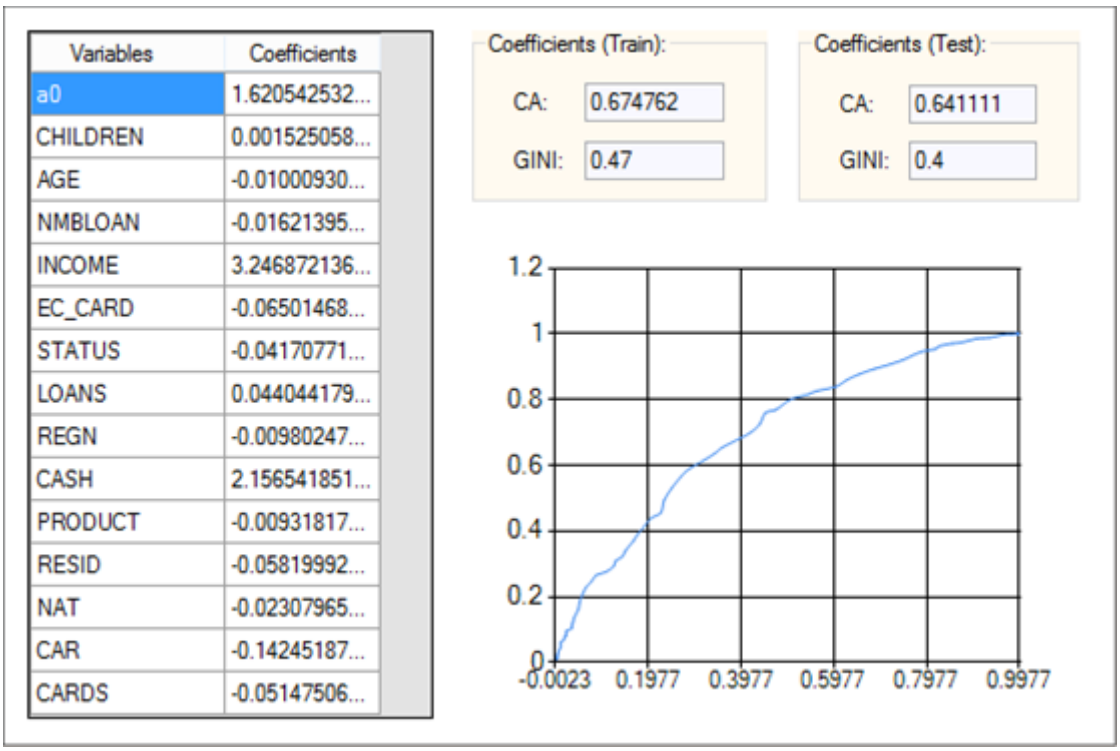

Рисунок 3.16 — Результати прогнозування з використанням WOE

Статистичні характеристики якості побудованих моделей було зведено до таблиці 3.2.

Таблиця 3.2 — Порівняльна таблиця характеристик якості скорингових моделей

	Навчальна вибірка (Train)		Тестова вибірка (Test)	
	Загальна точність (CA)	Індекс GINI	Загальна точність (CA)	Індекс GINI
Модель без використання WOE	0.658571	0.45	0.62	0.37
Модель з використанням WOE	0.674762	0.47	0.64111	0.4

Підсумовуючи результати з таблиці 3.2, можна зробити висновок, що побудована модель з використанням коефіцієнту WOE для трансформації категоріальних змінних в числові має кращі значення індексу GINI та загальної точності моделі на навчальній та тестовій вибірках.

Для порівняння якості побудованої скорингової моделі в програмному продукті Logist analytics з використанням методу логістичної регресії, було побудовані моделі у вигляді дерев рішень і скорингової карти в системі SAS Enterprise Miner [23].

Отримані моделі у вигляді скорингової карти та дерев рішень зображені на рисунках 3.17 та 3.18.

Scorecard

		Scorecard Points
Age	AGE< 23	-10
	23<= AGE< 28	1
	28<= AGE< 31	11
	31<= AGE< 46, _MISSING_	19
	46<= AGE	31
Credit Cards	CHEQUE CARD, MASTERCARD/EUROC, OTHER CREDIT CAR	32
	AMERICAN EXPRESS, NO CREDIT CARDS, VISA MYBANK, VISA OTHERS, _MISSING_, _UNKNOWN_	5
EC_card holders	0.00, _MISSING_, _UNKNOWN_	15
	1.00	6
Income	INCOME< 1000, _MISSING_	18
	1000<= INCOME< 1900	7
	1900<= INCOME< 2500	9
	2500<= INCOME< 3000	12
	3000<= INCOME	15

Рисунок 3.17 — Побудована скорингова карта в SAS Enterprise Miner

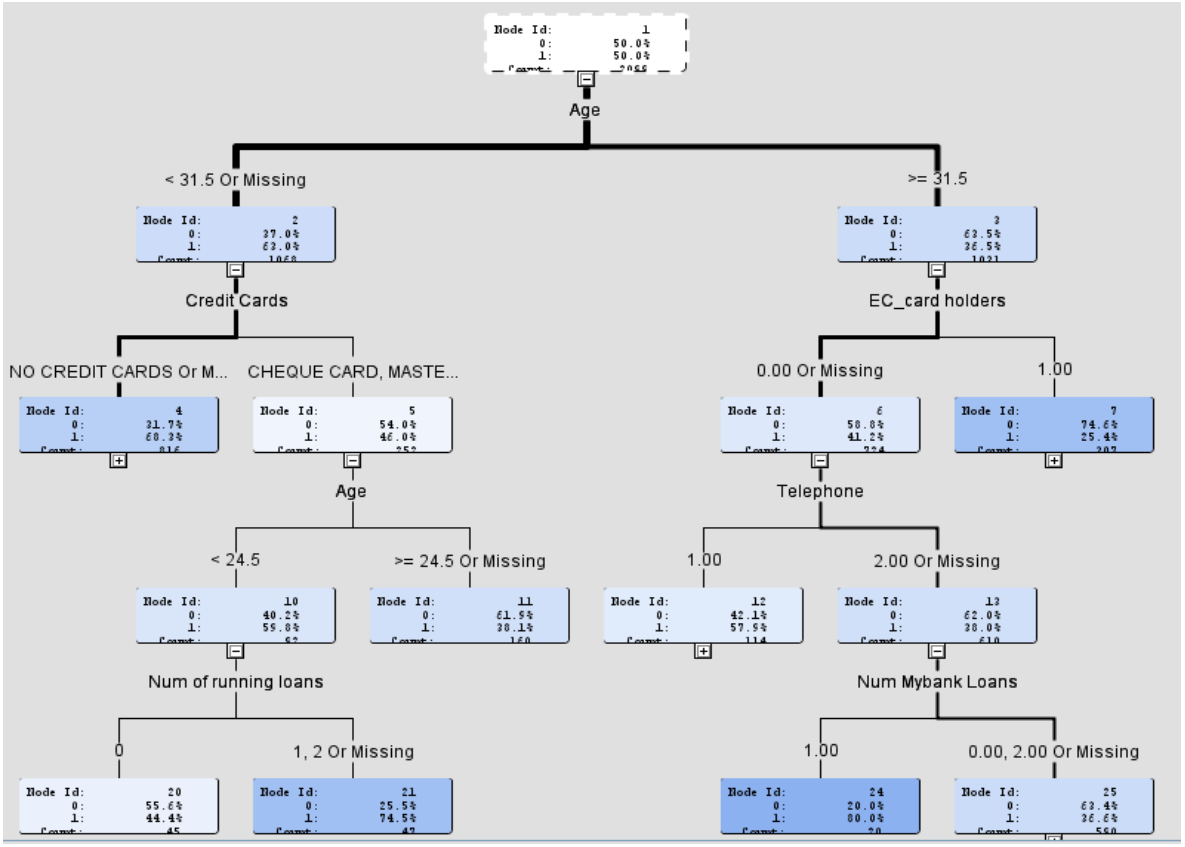

Рисунок 3.18 — Побудоване дерево рішень в SAS Enterprise Miner

Порівняння статистичних характеристик якості досліджуваних моделей наведено у таблиці 3.3.

Таблиця 3.3 – Порівняльна таблиця характеристик побудованих моделей

	Навчальна вибірка (Train)		Тестова вибірка (Test)	
	Загальна точність (CA)	Індекс GINI	Загальна точність (CA)	Індекс GINI
Логістична регресія (Logist analytics)	0.674762	0.47	0.641111	0.4
Дерева рішень (SAS Enterprise Miner)	0.686041	0.445	0.681465	0.403
Скорингова карта (SAS Enterprise Miner)	0.665079	0.476	0.622642	0.419

Найкращий результат за показником загальної точності моделі (CA) дав метод дерев рішень в системі SAS Enterprise Miner, на другому місці опинилась логістична регресія, реалізована у розробленому програмному продукті, значення загальної точності якої відрізняється лише на 0.01 від загальної точності дерев рішень. За індексом GINI найкращою виявилась скорингова карта, на другому місці – логістична регресія.

Висновки до розділу 3

У третьому розділі було описано спроектовану систему підтримки прийняття рішень для оцінки кредитоспроможності фізичних осіб. Дана система складається з таких структурних елементів: пристрої вводу-виводу, підсистема інтерфейсу користувача, підсистема зберігання інформації, підсистема обробки інформації, блок аналізу та прогнозування, і виведення результатів.

На основі запропонованої СППР в рамках магістерської дисертації було розроблено ПП Logist analytics. Програмний продукт дозволяє завантажувати

дані, проводити попередній аналіз та обробку даних, будувати прогнозуючі моделі, а також обчислювати статистичні характеристики якості побудованої моделі та зберігати результати прогнозування.

Визначено мінімальні технічні характеристики персонального комп'ютера для коректної та повноцінної роботи програмного забезпечення, а саме: тактова частота процесору, об'єм оперативної пам'яті, об'єм пам'яті на диску, операційна система, додаткове програмне забезпечення, що підтримує роботу розробленого програмного продукту, та периферійні пристрої необхідні для повноцінної роботи оператора.

Проведено детальний огляд інтерфейсу користувача. Розглянуто функціональні можливості програмного забезпечення та описану покрокову роботу Logist analytics з візуальним відображенням у вигляді рисунків робочого екрану програмного продукту. Побудовано дві скорингові моделі: з використанням коефіцієнту WOE для трансформації категоріальних змінних в числові, та з використанням порядкової нумерації. Порівнявши статистичні характеристики побудованих моделей можна побачити, що використання коефіцієнту WOE значно покращує прогнозуючу здатність моделі.

Було проведено порівняльний аналіз методу логістичної регресії з методами дерев рішень та скорингової карти, побудованими в системі SAS Enterprise Miner. За критерієм загальної точності моделі найкращий результат виявився у методу дерев рішень, а індекс Gini виявився найкращим у скорингової карти. Логістична регресія в Logist analytics показала середні результати за обома критеріями, що вказує на її хорошу прогнозуючу здатність.

4 РОЗРОБКА СТАРТАП-ПРОЕКТУ

На сьогоднішній день великої популярності набуває такий вид підприємництва як стартап. Стартап-проект – є комерційним проектом, який знаходиться в стані розробки, або нещодавно вийшов на ринок. Характерною особливістю стартапу, що відрізняє його від малого бізнесу, є оригінальність та інновації, він не може бути копією вже реалізованих ідей. При цьому проект не обов’язково повинен бути масштабного характеру, головне, щоб він був креативним, а його завдання – спрощувати людям будь-які дії в їх повсякденному житті.

Наразі, з появою Інтернету та сучасних технологій, стало простіше заходити на ринок, знаходити інвесторів та споживачів. З’явилося набагато більше можливостей для розвитку свого проекту за кордоном, ніж раніше. Проте розробка стартапу є досить ризикованим завданням. Не всім вдається довести свій стартап-проект до ринкового впровадження. За статистикою успіху досягає лише 10-20% від усіх стартап-проектів.

Запуск стартапу передбачає цілий ряд обов'язкових дій, в межах яких визначають ринкові перспективи стартапу, графік розробки, принципи організації виробництва, заходи з залучення інвесторів та аналіз ризиків.

4.1 Опис ідеї проекту

У таблиці 4.1 подано зміст ідеї стартап-проекту, можливі напрямки застосування та основні вигоди, що може отримати користувач товару. У таблиці 4.2 визначені сильні, слабкі та нейтральні сторони проекту.

Таблиця 4.1 — Опис ідеї стартап-проекту

Зміст ідеї	Напрямки застосування	Вигоди для користувача
Програмний продукт для прогнозування кредитоспроможності фізичних осіб на основі застосування теорії регресійного аналізу	Банківські установи	Дозволяє користувачам з різним рівнем підготовки проводити необхідну попередню обробку даних для побудови прогнозуючої моделі, будувати скорингову модель та одержувати прогнозні дані на основі побудованої моделі

Таблиця 4.2 — Порівняльна характеристика ідеї проекту

№ п/п	Техніко-економічні характеристики ідеї	Logist analytics	IBM SPSS Modeler	SAS Enterprise Miner
1	Ціна	Низька	Висока	Висока
2	Функціонал	Вузький	Широкий	Широкий

Отже, з табл. 4.2 можна визначити, що ціна є сильною характеристикою для потенційного товару, а функціонал, зважаючи на напрямки застосування товару, є нейтральною властивістю.

4.2 Технологічний аудит ідеї проекту

За результатами аналізу таблиці 4.3 можна зробити висновок про можливість технологічної реалізації проекту.

Таблиця 4.3 — Технологічна здійсненність ідеї проекту

№ п/п	Ідея проекту	Технології її Реалізації	Наявність технологій	Доступність технологій
1	Програмний продукт для прогнозування кредитоспроможності фізичних осіб на основі застосування теорії регресійного аналізу	Прогнозування на основі методу лінійної регресії	Наявна	Доступна
2		Прогнозування на основі методу логістичної регресії	Наявна	Доступна
Обрана технологія реалізації ідеї проекту: прогнозування на основі методу логістичної регресії				

4.3 Аналіз ринкових можливостей запуску стартап-проекту

Визначення ринкових можливостей, які можна використати під час ринкового впровадження проекту, та ринкових загроз, які можуть перешкодити реалізації проекту, дозволяє спланувати напрями розвитку проекту із урахуванням стану ринкового середовища, потреб потенційних клієнтів та пропозицій проектів-конкурентів.

Проведемо аналіз попиту: наявність попиту, обсяг, динаміка розвитку

ринку (табл. 4.4).

Таблиця 4.4 – Попередня характеристика потенційного ринку стартапу

№ п/п	Показники стану ринку (найменування)	Характеристика
1	Загальний обсяг продаж, грн/ум.од	100 000 ум.од
2	Динаміка ринку (якісна оцінка)	Зростає
3	Наявність обмежень для входу (вказати характер обмежень)	Немає
4	Специфічні вимоги до стандартизації та сертифікації	Немає
5	Середня норма рентабельності в галузі (або по ринку), %	75%

За результатами аналізу таблиці 4.4 можна зробити висновок, що ринок є привабливим для входження за попереднім оцінюванням.

Визначимо потенційні групи клієнтів, їх характеристики, та сформуємо орієнтовний перелік вимог до товару для кожної групи (табл. 4.5).

Таблиця 4.5 — Характеристика потенційних клієнтів стартап-проекту

№ п/п	Потреба, що формує ринок	Цільова аудиторія	Відмінності у поведінці груп клієнтів	Вимоги споживачів
1	Прийняття рішення щодо видачі кредитів фізичним особам	Банківські установи	Відмінність сфер діяльності клієнтів (кредитування фізичних осіб, юридичних осіб)	Висока точність прогнозування. Простий у використанні. Швидкодія при обробці значного об’єму інформації

Проведемо аналіз ринкового середовища: таблиці факторів, що сприяють ринковому впровадженню проекту, та факторів, що йому перешкоджають (табл. 4.6-4.7).

Таблиця 4.6 — Фактори загроз

Фактор	Зміст загрози	Можлива реакція компанії
Наявність великої конкуренції	Вихід на ринок великої компанії	Вихід з ринку. Обрати нову цільову аудиторію. Передбачити переваги продукту, щоб повідомити про них саме після виходу великої компанії на ринок
Зміна потреб користувачів	Користувачам необхідні рішення з іншим функціоналом	Передбачити можливість додавання нового функціоналу до продукту

Таблиця 4.7 — Фактори можливостей

Фактор	Зміст можливості	Можлива реакція компанії
Відсутність конкуренції	Відсутність аналогічних продуктів для користувача на вітчизняному ринку	Локалізація та адаптація сервісу для локальних груп. Адаптація до вітчизняних собливостей
Поява нових цільових груп Клієнтів	Потреба в аналогічному продукті в інших сферах діяльності	Адаптація продукту під нові сфери використання

Проведемо аналіз пропозиції: визначимо загальні риси конкуренції на ринку (табл. 4.8).

Таблиця 4.8 — Ступеневий аналіз конкуренції на ринку

Особливості конкурентного середовища	В чому проявляється дана характеристика	Вплив на діяльність підприємства (можливі дії компанії, щоб бути конкурентоспроможною)
1. Вказати тип конкуренції - монополістична	Існує декілька фірм-конкурентів	Підтримка якості продукту та постійні вдосконалення
2. За рівнем конкурентної боротьби - інтернаціональний	Фірми конкуренти з різних країн	Підтримувати продукт на національному ринку
3. За галузевою ознакою - внутрішньогалузева	Продукт використовується в одній галузі	Вдосконалювати продукт для застосування в інших галузях
4. Конкуренція за видами товарів: - товарно-родова	Присутня конкуренція з боку товарів-замінників	Розширювати функціонал продукту
5.За характером конкурентних переваг - нецінова	Вдосконалення якості продукції, технології виробництва, інновацій	Випускати нові товари, які принципово відрізняються від своїх попередників та представляють модернізований варіант старої моделі
6. За інтенсивністю – немарочна	Роль торгової марки незначна	Приділяти увагу якості продукту а не бренду компанії

Таблиця 4.9 — Обґрунтування факторів конкурентоспроможності

Фактор конкурентоспроможності	Обґрунтування (наведення чинників, що роблять фактор для порівняння конкурентних проектів вагомим)
Ціна	Більш доступна ціна збільшує кількість потенційних клієнтів
Функціонал	Функціонал направлений на предметну область
Зручний інтерфейс	Зручний інтерфейс робить продукт більш привабливим для клієнтів

За визначеними факторами конкурентоспроможності (табл. 4.9) проведемо аналіз сильних та слабких сторін стартап-проекту (табл. 4.10).

Таблиця 4.10 — Порівняльний аналіз сильних та слабких сторін «Logist analytics»

Фактор конкурентоспроможності	Бали 1-20	Рейтинг товарів-конкурентів у порівнянні з “Logist analytics”						
		–3	–2	–1	0	+1	+2	+3
Ціна	18		+					
Функціонал	10					+		
Зручний інтерфейс	12				+			

Складемо SWOT-аналіз (матриця аналізу сильних (Strength) та слабких (Weak) сторін, загроз (Troubles) та можливостей (Opportunities)) (табл. 4.11) на основі виділених ринкових загроз та можливостей, та сильних і слабких сторін (табл. 4.10).

Таблиця 4.11 – SWOT-аналіз стартап-проекту

Сильні сторони: ціна, зручний інтерфейс	Слабкі сторони: функціонал
Можливості: Низька конкуренція, поява нових потреб споживачів	Загрози: Висока конкуренція, невідповідність потребам споживачів

На основі SWOT-аналізу визначимо альтернативи ринкової поведінки (перелік заходів) для виведення стартап-проекту на ринок та орієнтовний оптимальний час їх ринкової реалізації з огляду на потенційні проекти конкурентів, що можуть бути виведені на ринок (табл. 4.12).

Таблиця 4.12 — Альтернативи ринкового впровадження стартап-проекту

Альтернатива (орієнтовний комплекс заходів) ринкової Поведінки	Ймовірність отримання ресурсів	Строки реалізації
Створення програмного забезпечення	80%	3 місяців
Створення веб-сервісу	60%	5 місяців

4.4 Розроблення ринкової стратегії проекту

Розроблення ринкової стратегії першим кроком передбачає визначення стратегії охоплення ринку: опис цільових груп потенційних споживачів (табл. 4.13).

Таблиця 4.13 – Вибір цільових груп потенційних споживачів

№ п/п	Опис профілю цільової групи клієнтів	Готовність споживачів сприйняти продукт	Орієнтовний попит в межах цільової групи (сегменту)	Інтенсивність конкуренції в сегменті	Простота входу у сегмент
1	Банки	Висока	Високий	Середня	Середня складність
2	Інші фінансові установи	Середня	Середній	Помірна	Висока складність

Для роботи в обраних сегментах ринку сформуємо базову стратегію розвитку (табл. 4.14).

Таблиця 4.14 — Визначення базової стратегії розвитку

Обрана альтернатива розвитку проекту	Стратегія охоплення ринку	Ключові конкуренто-спроможні позиції	Базова стратегія розвитку
Надання товару важливих з точки зору споживача відмітних властивостей, які роблять товар відмінним від товарів конкурентів	Визначити потреби кожної з цільових груп, розробити стратегії приваблення споживачів	Оперативне реагування на зміни в ринковому попиті, клієнто-орієнтованість, висока якість продукту	Стратегія диференціації

Наступним етапом є обрання вектору конкурентної поведінки (табл. 4.15).

Таблиця 4.15—Визначення базової стратегії конкурентної поведінки

Чи є проект «першопрохідцем» на ринку?	Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів?	Чи буде компанія копіювати основні характеристики товару конкурента, і які?	Стратегія конкурентної поведінки
Не є першопрохідцем	Шукати нових	Ні	Стратегія заняття конкурентної ніші

Сформуємо ринкову позицію, за якою споживачі мають ідентифікувати проект (табл. 4.16).

Таблиця 4.16 — Визначення стратегії позиціонування

Вимоги до товару цільової аудиторії	Базова стратегія розвитку	Ключові конкуренто-спроможні позиції власного проекту	Вибір асоціацій (комплексна позиція власного продукту)
Простий та зручний користувацький інтерфейс, надійність та безпека, швидкість роботи продукти	Стратегія диференціації	Позиція на основі порівняння продукту компанії з продуктами конкурентів. Відмінні особливості Споживачів	Автоматизація робочих процесів, зниження кредитних ризиків, зниження навантаження та часу

4.5. Розроблення маркетингової програми стартап-проекту

У нижченаведеній таблиці підсумуємо результати попереднього аналізу конкурентоспроможності товару.

Таблиця 4.17 — Визначення ключових переваг концепції потенційного товару

Потреба	Вигода, яку пропонує товар	Ключові переваги перед конкурентами
Автоматизація робочих процесів	Продукт автоматизує такі процеси, як обробка даних та прийняття рішення щодо видачі кредиту	Після впровадження продукту процес прийняття рішення щодо видачі кредиту стає автоматизований
Зменшення кредитних ризиків	Продукт зменшує кредитні ризики банківських установ	Висока точність прогнозування знижує кредитні ризики банківських Установ
Зниження навантаження та часу	Продукт знижує навантаження на персонал банківських установ та зменшує час видачі кредиту	Персоналу банків не потрібно самостійно аналізувати великий об’єм даних, що знижує навантаження на прискорює роботу

У ході дослідження також побудовано трирівневу маркетингову модель товару (табл. 4.18).

Таблиця 4.18 — Опис трьох рівнів моделі товару

Рівні товару	Сутність та складові
I. Товар за задумом	Програмний продукт для прогнозування кредитоспроможності фізичних осіб. Повинен бути зручним, швидким та безпечним
II. Товар у реальному виконанні	Властивості/характеристики
	1. Попередня обробка даних. 2.Побудова скорингової моделі 3. Прогнозування кредитоспроможності.
	Якість: проходження тестування
	Пакування: відсутнє
	Марка: “Logist analytics”
III. Товар із підкріпленням	До продажу: відсутнє
	Після продажу: навчання персоналу, супровід, технічна підтримка
Вихідний код програмного продукту є закритим, та не передається клієнтам і третім особам.	

Визначимо цінові межі, якими необхідно керуватись при встановленні ціни на товар (табл. 4.19).

Таблиця 4.19—Визначення меж встановлення ціни

Рівень цін на товари-замінники	Рівень цін на товари-аналоги	Рівень доходів цільової групи Споживачів	Верхня та нижня межі встановлення ціни на товар/послугу
2500$	2000$	Високий рівень доходів	Базова покупка та впровадження: нижня межа

Визначимо оптимальну систему збуту (табл. 4.20).

Таблиця 4.20 — Система збуту

Специфіка закупівельної поведінки цільових клієнтів	Функції збуту, які має виконувати постачальник товару	Глибина каналу збуту	Оптимальна система збуту
Цільові клієнти – банківські установи, які бажають впровадити у своїй роботі сучасні засоби,які допоможуть автоматизувати робочі процеси. Вони цікавляться інноваційними рішеннями, відвідують тематичні семінари та конференції	Формування попиту і стимулювання збуту. Встановлення контактів із споживачами. Просування маркетингової інформації	Нульова або однорівнева (сервіс безпосередньо продається споживачам та через посередників)	Прямий канал збуту до споживача, мінімізувати витрати на додаткові канали збуту

Розроблена концепція маркетингових комунікацій, що спирається на попередньо обрану основу для позиціонування, визначену специфіку поведінки клієнтів (табл. 4.21).

Таблиця 4.21 — Концепція маркетингових комунікацій

Специфіка поведінки цільових клієнтів	Канали комунікацій, якими користуються цільові Клієнти	Ключові позиції, обрані для позиціонування	Завдання рекламного повідомлення	Концепція рекламного звернення
Цільові клієнти – банківські установи, що займаються кредитуванням фізичних осіб, і бажають автоматизувати процесс видачі кредиту, та зменшити кількість неповернутих кредитів. Вони цікавляться інноваційними рішеннями, відвідують тематичні семінари та конференції	Конференції, форуми, новини у сфері інноваційних технологій, періодичні видання у професійних галузях	Позиція на основі порівняння продукту компанії з продуктами конкурентів. Відмінні особливості споживачів	1)інформувати про новий продукт та його переваги; 2)сформувати сприятливу думку; 3)сформувати образ марки виробника, збільшити потік покупців	Зменшуємо кредитні ризики. Прискорюємо та автоматизуємо процес видачі кредитів

Висновки до розділу 4

В даному розділі проведено аналіз створення та виведення на ринок стартап-проекту на основі програмного продукту, який було розроблено в рамках магістерської дисертації.

В межах цього аналізу було розроблено опис самої ідеї проекту, визначено загальні напрями використання товару, проаналізовано ринкові можливості щодо впровадження проекту, визначено відмінності від конкурентів та розроблено стратегію виходу на ринок.

Узагальнюючи проведений аналіз, можна зазначити, що є можливість ринкової комерціалізації проекту. Наявний попит, динаміка ринку зростає. З огляду на потенційні групи клієнтів, а саме банківські установи, та високий рівень конкурентоспроможності проекту, є достатні перспективи для впровадження стартапу. Отже, подальша імплементація проекту є доцільною.

ВИСНОВКИ

Дана робота присвячена аналізу, побудові та використанню прогнозуючої моделі для оцінювання кредитоспроможності позичальників і ризику банків при наданні кредитів. У зв'язку з нинішнім кризовим станом у банківській сфері постає нагальним застосування та розробка нових більш досконалих методів оцінювання кредитних ризиків і кредитоспроможності осіб.

У першому розділі магістерської роботи розглянуто понятійно-категоріальний апарат дослідження, а саме основний інструмент з мінімізації ризику кредитоспроможності позичальників — оцінки кредитоспроможності фізичних осіб. У цьому ж розділі поставлена задача дослідження у межах магістерської роботи.

У другому розділі проаналізовано існуючі математичні методи прогнозування кредитоспроможності фізичних осіб, а саме мережа Байєса, лінійна ймовірнісна модель, логістична регресія та скорингова карта.

У третьому розділі було розроблено програмний продукт Logist analytics з використанням технологій .Net у середовищі розробки Microsoft Visual Studio 2012. У системі було реалізовано метод логістичної регресії для прогнозування кредитоспроможності позичальників. Для знаходження оцінок параметрів регресійної моделі було використано метод максимальної правдоподібності з використанням методу градієнтного спуску. Також для якісної оцінки результатів побудови моделі логістичної регресії в системі Logist analytics, здійснено розробку скорингових карт в системі SAS Enterprise Miner на основі методів дерев рішень та логістичної регресії.

У четвертому розділі проведено аналіз створення та виведення на ринок стартап-проекту на основі програмного продукту Logist analytics. В межах цього аналізу було здійснено опис самої ідеї проекту, визначено загальні

напрями використання товару, встановлено ринкові можливості щодо впровадження проекту та розроблено стратегію виходу на ринок.

Результати магістерської дисертації:

- розроблено структуру системи підтримки прийняття рішень для оцінювання кредитоспроможності фізичних осіб;

- реалізовано програмний продукт для аналізу та обробки даних, побудови скорингової моделі на основі логістичної регресії для прогнозування кредитоспроможності;

- застосовано метод максимальної правдоподібності з використанням методу градієнтного спуску для побудови моделі логістичної регресії;

- результати прогнозування кредитоспроможності в системі ПП у порівнянні з іншими методами прогнозування, реалізованими в комерційних системах.

Наступними напрямками роботи можуть бути вирішення питання, що стосуються:

- удосконалення розробленого методу побудови скорингової моделі;

- розробки нових підходів щодо визначення ступеня значимості регресорів;

- реалізації методів інтелектуального аналізу даних іншого типу, наприклад, нейронних мереж, методу групового врахування аргументів, методу опорних векторів.

Розроблений програмний продукт показав прийнятні результати, що підтверджує раціональність використання обраного методу.

ПЕРЕЛІК ПОСИЛАНЬ

1. Allison P.D. Logistic regression using the SAS system: theory and application. Cary: SAS Institute Inc., 1999. 287 p.
2. Anderson B. S. Developing Credit Scorecards Using SAS Credit Scoring for Enterprise Miner 5.3. Cary: SAS Institute Inc, 2009. 41 p.
3. Anderson B.S., Thompson R.W. Developing Credit Scorecards Using SAS Credit Scoring for Enterprise Miner 5.3. Cary: SAS Institute Inc, 2009. 41 p.
4. Aven T. Foundations of Risk Analysis: A Knowledge and Decision-Oriented Perspective. New York: John Wiley & Sons, Ltd., 2003. 198 p.
5. Bielecki T.R., Rutkowsky M. Credit Risk: Modeling, Valuation, Hedging. Berlin: Springer, 2002. 500 p.
6. H. Van Gruening, S. B. Bratanovic. Analyzing and Managing Banking Risks. Washington: The World Bank, 2003. 386 p.
   http://www.nbuv.gov.ua/old_jrn/Soc_Gum/Uproz/2012_14/u1214 mur.pdf .
7. Liu Y. New issues in credit scoring application / Institut für Wirtschaftsinformatik: Arbeitsberich, 2001. № 16.
8. Mok Jie-Men. Reject Inference in Credit Scoring. Amsterdam: BMI paper, 2009. 38 p.
   P. 149–172.
9. Thomas L. C. A survey of credit and behavioral scoring: forecasting financial risk of lending to consumers. International Journal of Forecasting, 2000. Vol. 16.
10. Азаренкова Г.М. Рейтингове оцінювання як метод визначення кредитоспроможності позичальників. *Вісник університету банківської справи Національного банку України* : зб. наук. праць, 2011. № 1(10). 219 с.

11. Бідюк П.І., Коршевнюк Л.О. Проектування комп'ютерних інформаційних систем підтримки прийняття рішень: Навчальний посібник. Київ: ННК «ІПСА» НТУУ «КПІ», 2010. 340 с.
12. Боровиков В. Искусство анализа даних на компьютере. М.: 2003. 688 с.
13. Васюренко О.В. Банківські операції : навч. посібн. Вид. 6-т. – К. : Вид-во "Знання", 2008. 318 с.
14. Вдовенко Л. О. Економічна сутність та значення кредитоспроможності підприємств. Облік і фінанси, 2012. №1. С. 108-111. Режим доступу: http://nbuv.gov.ua/UJRN/Oif_apk_2012_1_23 .
15. Згуровський М. З., Бідюк П. І., Терентьєв О. М., Просянкіна-Жарова Т. І. Байєсівські мережі в системах підтримки прийняття рішень: навч. посіб. / Київ, 2015. 300 с.
16. Кабак А.Ф. Економетрія. Одеса, 2003. 561 с.
17. Камінський А. Б. Моделювання фінансових ризиків : монографія/ Київ: Видавничо- поліграфічний центр «Київський університет», 2006. 306 с.
18. Корольова О. Методичний інструментарій оцінки кредитоспроможності підприємства. Економічний аналіз, 2009. Вип. 4. С. 238-240.
19. Коршевнюк Л. А. Проблема поддержки принятия решений при управлении бизнес-процессами на предприятиях/ ред. Коршевнюк Л.А., Бидюк П. И..
20. Крістіогло Г. М. Формування ринкових відносин в Україні. Крістіогло Г. М. Використання скорингових моделей в умовах невизначеності та ризику споживчого кредитування., 2007. № 7 (74). С. 86–90.
21. Кузнєцова Н. В. Порівняльний аналіз характеристик моделей оцінювання ризиків кредитування / ред. Н. В. Кузнєцова, П. І. Бідюк / *Наукові вісті НТУУ «КПІ»*, 2010. №1. С. 42-53.
22. Кузнєцова Н.В. Системний підхід до аналізу кредитних ризиків з використанням мереж Байєса / Н.В. Кузнєцова, П.І. Бідюк / *Наук. вісті НТУУ «КПІ»*, 2008. № 3. С. 11–24.

23. Муравйова М.Ю. Шляхи вдосконалення оцінки кредитоспроможності позичальників банками України. Управління розвитком, 2012. №14 (135). Режим доступу:

24. Новак Е. Введение в методы эконометрики. М.: 2004. 245 с.

25. Носко В.П. Эконометрика для начинающих. М., 2005. 379 с.

26. Отонець К. Практичні аспекти застосування скорингу для оцінки кредитного ризику. Фінансовий ринок України, 2006. № 4. С. 14–17.

27. Самойлова С. С. Скоринговые модели оценки кредитного риска Социально-экономические явления и процессы, 2014. №3 (61). Режим доступа: http://cyberleninka.ru/article/n/skoringovye-modeli-otsenki-reditnogo-riska .

28. Сиддики Н. Скоринговые карты для оценки кредитных рисков. Разработка и внедрение интеллектуальных методов кредитного скоринга / пер. с англ. Е. Ильичева. М.: Манн, Иванов и Фербер, 2014. 268 с.

Системні технології. 2000, №3(11). С. 40-51.

29. Ситник В. Ф. Системи підтримки прийняття рішень: навч. посіб. К.: КНЕУ, 2004. 614 с.

30. Сорокин М. "ЦЕНЗОР" – новая система оценки кредитоспособности заемщика. Экономика Украины. 1999. № 4. С. 24-32.

ДОДАТОК А ТАБЛИЦІ СТАТИСТИЧНИХ ДАНИХ

Таблиця А. 1—Вихідні статистичні дані щодо перших 9 показників

TITLE	CHILDREN	AGE	NMBLOAN	FLOAN	INCOME	ECARD	STATUS	LOANS
R	0	46	0	0	0	1	V	0
H	4	34	1	1	3200	0	V	2
H	3	31	1	1	3300	0	V	3
R	0	39	0	0	1500	0	W	1
H	3	32	2	1	0	1	V	1
H	0	23	2	1	0	1	U	1
R	0	42	0	0	1900	0	V	0
H	2	35	0	0	0	1	V	2
H	1	26	0	0	1700	1	V	0
H	1	24	2	1	3400	0	U	2
R	0	20	0	0	0	1	U	0
H	1	44	0	1	0	1	G	2
H	0	23	0	0	1900	0	U	1
H	1	44	0	0	0	1	V	0
H	6	36	2	1	3500	0	V	0
H	1	31	2	1	3200	1	V	6
H	0	56	2	1	2200	0	U	0
R	0	42	2	1	1600	0	U	1
R	0	42	2	1	1500	0	U	1
H	2	35	2	1	3500	0	V	3
R	2	30	2	1	0	1	G	1
H	0	21	2	0	2700	0	U	1
H	0	57	2	1	2000	0	V	2
H	0	29	2	1	2600	0	V	2
H	3	33	1	1	3200	0	V	2
H	0	26	2	1	2300	0	U	2
R	0	31	2	1	2100	0	G	1
R	0	27	2	1	2200	0	V	1

Таблиця А.2 – Вихідні статистичні дані щодо останніх 7 показників

REGN	CASH	PRODUCT	RESID	CAR	CARDS	GB
0	2000	Radio, TV, Hifi	Lease	Car	Cheque card	0
4	6000	None	Owner	Car	no credit cards	1
4	0	None	Lease	Car	no credit cards	1
0	2500	Furniture,Carp	Lease	Without	no credit cards	1
0	2500	Furniture,Carp	Lease	Car	Cheque card	0
0	700	Furniture,Carp	Lease	Car	Cheque card	1
0	9000	Furniture,Carp	Lease	Car	no credit cards	0
0	1700	Furniture,Carp	Lease	Car	Cheque card	0
0	800	Radio, TV, Hifi	Lease	Car	Cheque card	1
0	1000	Radio, TV, Hifi	Lease	Car	no credit cards	1
0	4000	Furniture,Carp	Lease	Without	Cheque card	0
0	800	Radio, TV, Hifi	Lease	Car	Cheque card	0
0	600	Dept. Store,Mail	Lease	Car	no credit cards	1
0	6000	Furniture,Carp	Lease	Car	Cheque card	1
0	1000	Dept. Store,Mail	Lease	Car	no credit cards	0
6	600	Radio, TV, Hifi	Lease	Car	Cheque card	0
4	3000	Radio, TV, Hifi	Lease	Without	no credit cards	1
3	700	Dept. Store,Mail	Lease	Without	no credit cards	1
3	1800	Radio, TV, Hifi	Lease	Without	no credit cards	1
4	1200	Dept. Store,Mail	Lease	Without	no credit cards	0
9	1100	Furniture,Carp	Lease	Car	Cheque card	1
8	700	Radio, TV, Hifi	Lease	Car	no credit cards	1
4	600	Radio, TV, Hifi	Lease	Car	no credit cards	0
5	1000	Radio, TV, Hifi	Lease	Car	no credit cards	0
5	2500	Radio, TV, Hifi	Lease	Car	no credit cards	1
2	5000	Radio, TV, Hifi	Lease	Without	no credit cards	1
5	3000	Furniture,Carp	Lease	Car	no credit cards	0
5	2500	Radio, TV, Hifi	Lease	Car	no credit cards	1
4	4000	Furniture,Carpet	Lease	Car	no credit cards	0
4	2500	Radio, TV, Hifi	Lease	Without	no credit cards	1

ДОДАТОК Б. ЛІСТИНГ ПРОГРАМНОГО ПРОДУКТУ

Б 1. Програмний код методу максимальної правдоподібності

```
class MMP
{
public static double f(double[] x, double[] Teta)
{
var result = (double)(1 / (1 + Math.Exp(-1 * Matrix.MultT1(Teta, x)))); return result;
}
public static double[] grad(double[,] X, double[] y, double[] Teta)
{
double[] sum = new double[Teta.Length]; for (int i = 0; i < sum.Length; ++i)
{
sum[i] = 0;
}

for (int i = 0; i < y.Length; ++i)
{
double[] x = Matrix.GetRow(X, i);
sum = Matrix.Sum(sum, Matrix.MultNum(x, (y[i] - f(x, Teta))));
}
return sum;
}

public static double Error(double[,] X, double[] y, double[] Teta)
{
double sumSquaredError = 0;
for (int i = 0; i < y.Length; ++i) // each data
{
double[] x = Matrix.GetRow(X, i); double computed = f(x, Teta); double desired = y[i];
sumSquaredError += (computed - desired) * (computed - desired);
}
return sumSquaredError / y.Length;
```

Б.2 Програмний код матричних процедур

```
class Matrix
{
public static double[,] Mult(double[,] matr_1, double[,] matr_2)
{
if (matr_1.GetLength(1) != matr_2.GetLength(0)) throw new Exception(“Error”);
double[,] result = new double[matr_1.GetLength(0), matr_2.GetLength(1)]; // 0-rows , 1-column for (int i = 0; i < matr_1.GetLength(0); i++)
{
for (int j = 0; j < matr_2.GetLength(1); j++)
{
for (int k = 0; k < matr_2.GetLength(0); k++) result[i, j] += matr_1[i, k] * matr_2[k, j];
}
}
return result;
}

public static double[,] MultT(double[,] Tmatr_1, double[,] Tmatr_2)
{
if (Tmatr_1.GetLength(0) != Tmatr_2.GetLength(1)) throw new Exception(“Error”); double[,] matr_2 = Tran(Tmatr_2);
double[,] matr_1 = Tran(Tmatr_1);
double[,] result = new double[matr_1.GetLength(0), matr_2.GetLength(1)]; // 0-rows , 1-column for (int i = 0; i < matr_1.GetLength(0); i++)
{
for (int j = 0; j < matr_2.GetLength(1); j++)
{
for (int k = 0; k < matr_2.GetLength(0); k++) result[i, j] += matr_1[i, k] * matr_2[k, j];
}
}
return result;
}
```

```
public static double[] Mult(double[,] matr, double[] vec)
{
if (matr.GetLength(1) != vec.Length) throw new Exception("Матрицы нельзя перемножить"); double[] result = new double[matr.GetLength(0)];
for (int i = 0; i < matr.GetLength(0); i++)
{
for (int j = 0; j < vec.Length; j++) result[i] += matr[i, j] * vec[j];
}
return result;
}

public static double[,] Tran(double[,] matr)
{
double[,] result = new double[matr.GetLength(1), matr.GetLength(0)]; for (int i = 0; i < matr.GetLength(0); i++)
for (int j = 0; j < matr.GetLength(1); j++) result[j, i] = matr[i, j];
return result;
}

public static double[] GetColumn(double[,] matr, int index)
{
if (matr.GetLength(1) <= index) throw new Exception("Error index > demension");

double[] result = new double[matr.GetLength(0)]; for (int i = 0; i < matr.GetLength(0); i++)
result[i] = matr[i, index]; return result;
}

public static double[] GetRow(double[,] matr, int index)
{
if (matr.GetLength(0) <= index) throw new Exception("Error index > demension"); double[] result = new double[matr.GetLength(1)];
for (int i = 0; i < matr.GetLength(1); i++) result[i] = matr[index, i];
return result;
}
```

```
public static double[,] MultNum(double[,] matr, double number)
{
double[,] result = (double[,])matr.Clone(); for (int i = 0; i < result.GetLength(0); i++)
for (int j = 0; j < result.GetLength(1); j++) result[i, j] *= number;
return result;
}

public static double[] MultNum(double[] vec, double number)
{
double[] result = (double[])vec.Clone(); for (int i = 0; i < result.GetLength(0); i++)
result[i] *= number; return result;
}

public static double[,] Sum(double[,] matr_1, double[,] matr_2)
{
if ((matr_1.GetLength(0) != matr_2.GetLength(0)) || (matr_1.GetLength(1) != matr_2.GetLength(1))) throw new Exception("Error");
double[,] result = (double[,])matr_1.Clone(); for (int i = 0; i < result.GetLength(0); i++)
for (int j = 0; j < result.GetLength(1); j++) result[i, j] += matr_2[i, j];
return result;
}
public static double[] Sum(double[] vec_1, double[] vec_2)
{
if (vec_1.Length != vec_2.Length) throw new Exception("Error"); double[] result = (double[])vec_1.Clone();
for (int i = 0; i < result.Length; i++) result[i] += vec_2[i];
return result;
}
public static double[] Sum(double[] vec, double num)
{
double[] result = (double[])vec.Clone(); for (int i = 0; i < result.Length; i++)
result[i] += num; return result;
}
```

Б.3 Програмний код дискретизації змінних

```
public partial class Discret
{
private double[,] Uniform_count(double[] X, int n)
{
double[,] Categ_Uniform = new double[n, 2]; int N = X.Length;
double[] X_sort = new double[N]; for (int i = 0; i < N; i++)
X_sort[i] = X[i]; Array.Sort(X_sort);
int perc = N / n;
int ost = N - perc * n; int k = 0;
Categ_Uniform[0, 0] = Double.MinValue; Categ_Uniform[n - 1, 1] = Double.MaxValue;
for (int i = 0; i < n; i++)
{
int num = 1; if (i != 0)
Categ_Uniform[i, 0] = X_sort[k]; if (i != n - 1)
{
Categ_Uniform[i, 1] = X_sort[k + 1]; while (num < perc)
{
num++; k++;
Categ_Uniform[i, 1] = X_sort[k + 1];
}
k++;
if (ost > 0)
{
Categ_Uniform[i, 1] = X_sort[k + 1]; k++;
ost--;
}
while (X_sort[k - 1] == X_sort[k])
{
Categ_Uniform[i, 1] = X_sort[k + 1]; k++;
}
```

```
perc = (N - k) / (n - (i + 1));
ost = (N - k) - perc * (n - (i + 1));
}
}
return Categ_Uniform;
}

private double[,] Uniform_width(double[] X, int n)
{
double[,] Categ_Uniform = new double[n, 2]; double perc = (X.Max() - X.Min()) / n; double
temp = X.Min() + perc;
for (int i = 0; i < n; i++)
{
if ((i != 0) || (i != n - 1))
{

Categ_Uniform[i, 0] = temp - perc; Categ_Uniform[i, 1] = temp;
}
else
{
if (i == 0)
Categ_Uniform[0, 0] = Double.MinValue; else
Categ_Uniform[n - 1, 1] = Double.MaxValue;
}
temp += perc;
}
return Categ_Uniform;
}
}
```

Б.4 Програмний код перетворення категоріальних змінних в числові

```
private double[] ConvertWOE(List<string> list, double[] Y)
{
double[] Cat = new double[list.Count]; string[] Ar = new string[list.Count]; for (int i = 0; i < list.Count(); i++)
Ar[i] = list[i];
list.Sort();
list = list.Distinct().ToList();
double[] woe = new double[list.Count]; double p;
for (int i = 0; i < list.Count; i++)
{
int n = 0, n_y = 0;
for (int j = 0; j < Ar.Length; j++)
{
if (Ar[j] == list[i])
{
if (Y[j] == 0)
n_y++; n++;
}
}
p = (double)n_y / n;
woe[i] = Math.Log((double)p / (1 - p));
}
quickSort(list, woe, 0, woe.Length - 1); for (int i = 0; i < Ar.Length; i++)
{
for (int j = 0; j < list.Count; j++) if (Ar[i] == list[j])
Cat[i] = j+1;
}
return Cat;
}

private double[] ConvertCat(List<string> list)
{
double[] Cat = new double[list.Count]; string[] Ar = new string[list.Count]; for (int i = 0; i < list.Count(); i++)
Ar[i] = list[i];
```

```
list.Sort();
list = list.Distinct().ToList();
for (int i = 0; i < Ar.Length; i++)
{
for (int j = 0; j < list.Count; j++) if (Ar[i] == list[j])
Cat[i] = j + 1;
}
return Cat;
}
```

Б.5 Програмний код розрахунку статистичних коефіцієнтів якості

```
public double CA(double[,] X, double[] Y, double[] Teta)
{
int n = 0;
double[] res = Matrix.Mult(X, Teta); for (int i = 0; i < res.Length; i++)
{
if (res[i] >= 0.5) res[i] = 1; elseres[i] = 0;
if (res[i] == Y[i]) n++;
}
return (double)n / (double)res.Length;
}

public double Gini(double[,] X, double[] Y, double[] Teta, ref double[] Se, ref double[] Sp)
{
double[] cutoff = new double[100]; double[] res = Matrix.Mult(X, Teta);
double step = (double)(res.Max() - res.Min()) / (double)100; double temp = 0;
List<string> str = new List<string>(); for (int i = 0; i < cutoff.Length; i++)
{
cutoff[i] = (double)res.Min() + (double)temp; temp += step;
}
```

```
for (int k = 0; k < cutoff.Length; k++)
{
int TP = 0; int TN = 0; int FP = 0; int FN = 0; int res_temp = 0; for (int i = 0; i < res.Length; i++)
{
if (res[i] >= cutoff[k])res_temp = 1; elseres_temp = 0;
if (res_temp == Y[i])
{
if (Y[i] == 0) TP++;
elseTN++;
}
else
{
if (Y[i] == 0) FN++;
elseFP++;
}
}
Se[k] = Math.Round((double)TP / (double)((double)TP + (double)FN), 4); Sp[k] = Math.Round(1 - (double)TN / (double)((double)TN + (double)FP), 4);
}

}
double Square(double[] y, double[] x)
{
double res = 0;
for (int i = 0; i < y.Length - 1; ++i)
res += (double) (y[i] + y[i + 1]) * (x[i + 1] - x[i]) / (double) 2; return res;
}
```

Б.6 Програмний код розрахунку коефіцієнту інформаційного значення

```
double IV(List<string> list, double[] Y)
```

```
{
double iv = 0;
string[] Ar = new string[list.Count]; for (int i = 0; i < list.Count(); i++)
Ar[i] = list[i];
list.Sort();
list = list.Distinct().ToList();
for (int i = 0; i < list.Count; i++)
{
int n = 0, n_y = 0; double woe = 0; double p = 0;
for (int j = 0; j < Ar.Length; j++)
{
if (Ar[j] == list[i])
{
if (Y[j] == 0)
n_y++; n++;
}
}
p = (double)n_y / n;
woe = Math.Log((double)p / (1 - p)); if (double.IsInfinity(woe))
woe = 0;
iv += (p - (1 - p)) * woe;
}
return iv;
}
```